NORTH DAKOTA DEPARTMENT OF
PUBLIC INSTRUCTION

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

2020–2021

Volume 2
Test Development

ACKNOWLEDGMENTS

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

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

TABLE OF CONTENTS

LIST OF TABLES

LIST OF APPENDICES

1. INTRODUCTION

The Independent College and Career Readiness (ICCR) English language arts (ELA) and mathematics item bank is written to measure college and career readiness standards as embodied in the College- and Career-Ready Standards. The bank is designed to measure the full breadth and depth of the standards and cover a range of difficulty that matches the distribution of student performance in each grade and subject. The item bank is designed primarily for accountability assessments.

Items were developed for all reading and writing standards and a subset of the speaking and listening standards. The speaking and listening standards that are not covered in the bank include SL.1, SL.4, SL.5, and SL.6, as most states choose not to measure these standards on their accountability assessments.

All items were developed to meet detailed specifications that identified how items measuring each standard should do so. ICCR item specifications were developed as a joint partnership with the state of Utah. It quickly grew into a unique product with specifications that meet the needs of multiple states. CAI notes that the Utah Student Assessment of Growth and Excellence (SAGE), the first state assessment developed using ICCR items, received the approval of peer reviewers, validating the quality and alignment of the specifications to the College- and Career-Ready Standards. Over time, the specifications have been updated to incorporate an expanding set of potential interactions and item types. A growing pool of states adopted ICCR as a component of their item pool or, in some cases, the entire basis of their tests. As we describe, each ICCR item has been through a series of stakeholder reviews in one or more participating states. This information regarding the item bank remains consistent, but Sections 3 and 4 are updated annually to reflect item bank growth and observed test administration match to the blueprints.

1.1 CLAIM STRUCTURE

The item bank is designed to measure college and career readiness and can support tests that claim that students in grades 3–11 demonstrate progress toward college and career readiness in mathematics and ELA.

Within ELA, the items are designed to support the following claims about proficient students:

- Students can read closely and analytically to comprehend a range of increasingly complex literary and informational texts.
- Students can write well-structured, focused texts for a variety of purposes, analytically integrating information from multiple sources.
- Students know and can apply the rules of standard, written English.

In mathematics, tests built from the ICCR item bank can support claims such as the following: *Proficient students in Grade 7 can use procedures involving rational numbers to solve problems, model real-world phenomena, and reason mathematically*. The specific classes of procedure

vary by grade level and are summarized in Exhibit A. While grades 3–8 are ICCR generalizable, grade 10 is specific to North Dakota.

Exhibit A: ICCR Mathematics Procedural Categories Forming the Basis of Reporting Categories in Each Grade

Grade(s)	Classes of Procedures				
3, 4, 5	Operations and Algebraic Thinking	Number and Operations in Base Ten	Number and Operations in Fractions	Measurement, Data, and Geometry	—
6, 7	Expressions and Equations	Ratios and Proportional Relationships	Number Systems	Geometry	Statistics and Probability
8	Expressions and Equations	Number Systems	Functions	Geometry	Statistics and Probability
10	Algebra	Functions	Geometry	Statistics and Probability	—

1.2 Underlying Principles Guiding Development

The ICCR item bank was established using a highly structured, evidence-centered design. The process began with detailed item specifications. These specifications, discussed in a later section, described the interaction types that could be used, provided guidelines for targeting the appropriate cognitive engagement, offered suggestions for controlling item difficulty, and offered sample items.

Items were written with the goal that virtually every item would be accessible to all students, either by itself or in conjunction with accessibility tools, such as text-to-speech, translations, or assistive technologies. This goal is supported by the delivery of the items on CAI's test delivery platform, which has received an internationally recognized accessibility standard known as WCAG 2.0 AA certification. The test delivery platform offers a wide array of accessibility tools and is compatible with most assistive technologies.

Item development supported the goal of creating high-quality items through rigorous development processes managed and tracked by a content development platform that ensures that every item flows through the correct sequence of reviews and captures every comment and change to each item.

We sought to ensure that the items were measuring the standards in a fair and meaningful way by engaging educators and other stakeholders at each step of the process. Educators evaluated the alignment of items to the standards and offered guidance and suggestions for improvement. They participated in the review of items for fairness and sensitivity. Following the field testing of items,

educators engaged in *rubric validation*, a process that refines rule-based rubrics upon review of student responses.

In coordinating among states, educators in multiple states would frequently review the same items. In general, one state was assigned rights to modify the items, and other states were offered the modified items on an accept-reject basis.

Combined, these principles and the processes that support them have led to an item bank that measures the standards with fidelity and does so in a way that minimizes construct-irrelevant variance and barriers to access. The details of these processes follow.

1.3 Organization of This Volume

This volume is organized into three sections:

1. An overview of the item development process that supports the validity of the claims that ICCR tests are designed to support
2. An overview of the item pool, the types of assessments the pool is designed to support, and methods for refreshing the pool
3. A description of test construction for the North Dakota State Assessment (NDSA) for ELA and mathematics, including the blueprint design and test construction

2. Item Development Process That Supports Validity of Claims

2.1 Overview

CAI developed the ICCR ELA and mathematics item banks using a rigorous, structured process that engaged stakeholders at critical junctures. This process was managed by CAI's Item Tracking System (ITS), which is an auditable content development tool that enforces rigorous workflow and captures every change to, and comment about, each item. Reviewers, including internal CAI reviewers or stakeholders in committee meetings, can review items in ITS as they will appear to the student, with all accessibility features and tools.

The process begins with the definition of passage and item specifications, and continues with

- selection and training of item writers;
- writing and internal review of items;
- review by state personnel and stakeholder committees;
- markup for translation and accessibility features;
- field testing; and
- post field-test reviews.

Each of these steps has a role in ensuring that the items can support the claims that will be based on them. Exhibit B describes how each step contributes to these goals. Each step in the process is discussed in more detail.

Exhibit B: Summary of How Each Step of Development Supports the Validity of Claims

	Supports Alignment to the Standards	Reduces Construct-Irrelevant Variance Through Universal Design	Expands Access Through Linguistic and Other Supports
Passage and item specifications	Specifies item types, content limits, and guidelines for meeting Depth of Knowledge (DOK) requirements and adjusting difficulty.	Avoids the use of any item types with accessibility constraints and provides language guidelines. Allows for multiple response modes to accommodate different styles.	
Selection and training of item writers	Ensures that item writers have the background to understand the standards and specifications. Teaches item writers about selection of item types for measurement and accessibility.	Training in language accessibility, bias, and sensitivity helps item writers avoid unnecessary barriers.	
Writing and internal review of items	Checks content and DOK alignment and evaluates and improves overall quality.	Eliminates editorial issues, and flags and removes bias and accessibility issues.	
Markup for translation and accessibility features		Adds universal features, such as text-to-speech for mathematics, that reduce barriers.	Adds text-to-speech, braille, ASL, translations, and glossaries.
Review by state personnel and stakeholder committees	Checks content and DOK alignment; evaluates and improves overall quality.	Flags sensitivity issues.	
Field testing	Provides statistical check on quality and flags issues.	Flags items that appear to function differently for subsequent review for issues.	May reveal usability or implementation issues with markup.
Post field-test reviews	Final, more focused check on flagged items. Rubric validation and rangefinding ensure that scoring reflects standards and expectations.	Final, focused review on items flagged for differential item function.	

2.2 Passage and Item Specifications

Items and passage specifications were initially developed in collaboration between content experts in the Utah State Board of Education and CAI content experts in respect to College- and Career-Ready Standards. The specifications were used to develop both the SAGE pool and the ICCR pool, but align to nationally recognized standards. North Dakota state standards are similar to College- and Career-Ready Standards; therefore, CAI used the ICCR items to measure North Dakota

standards. Over time, the specifications have been expanded to reflect continuous improvement and the availability of new interaction types.

2.2.1 Passage Specifications

ELA development begins with passage specifications. Detailed passage specifications ensure that all passages align to the correct grade level and provide sufficient complexity for close analytical reading. These specifications augment, rather than replace, quantitative syntactic measures, such as Lexiles. The qualities called out in the specifications are derived from the Common Core State Standards (CCSS) ELA standards and accompanying material.

Exhibit C provides a sample passage specification.

Exhibit C: Sample Passage Specifications

Difficulty Factor	Passage Metric Description	Grade Level Detail (Sample for Grade 6)	Research-Based Evidence
Levels of Meaning in Literature	1. Single, concrete interpretation with few generalizations necessary 2. Some themes not explicitly stated 3. Multiple, successively abstract or general levels of meaning; key theme or themes implied	1. a. The passage has a single, concrete meaning conveyed through dialogue or narration. b. The main idea or theme is explicitly stated and clearly supported with supplementary details or quotes. c. Relationships between related concepts are clearly linked and defined. d. Characters and their motivations are explicitly defined in the passage. e. Setting is used as an aesthetic enhancement, not as a way to convey meaning. 2. a. The main idea or theme of the passage may be either explicitly or implicitly stated, but multiple connections must be made to understand the full impact. b. Actions of multiple characters are central to the theme and/or plot. Relationships between characters and characters' motivations require interpretation. c. Mood, setting, and tone may be easily identified but do not heavily	Research shows that concrete passages are more comprehensible and easier to recall than abstract passages (Sadoski, Goetz, & Fritz, 1993). Comprehension for concrete passages also increases in relation to how easily the reader can imagine the contents of the text (Riding & Taylor, 1976). Characterization, in particular, plays a role in a text's difficulty. When a character's actions are clearly linked to the character's emotional state, the text is much more readily comprehensible (Gillioz, Gygax, & Tapiero, 2012). Similarly, readers draw inferences from descriptions of a character's actions and stated preferences (i.e., descriptions of specific traits as being either positive

Difficulty Factor	Passage Metric Description	Grade Level Detail (Sample for Grade 6)	Research-Based Evidence
		influence the overall meaning or theme. 3. a. The passage contains several ideas and/or themes, both explicitly stated and implied. b. The reader must draw inferences about meaning from different elements of the passage, including character(s), setting, plot, dialogue, structure, and/or tone. c. Characters' motivations and characteristics are strongly implied through clear action or dialogue. d. Mood, setting, and tone may be subtle and have a greater impact on the overall meaning.	or negative) (Mensink & Rapp, 2011). However, when a character exhibits behavior that is inconsistent with a perceived trait, the characterization takes longer for readers to process and comprehend (Sparks & Rapp, 2011). An increase in dialogue between characters has a similar effect, as tested readers' response times to items about dialogue scenes were slower than for nondialogue scenes (Long & De Ley, 2000). Beyond-text inferences involving aspects of stories such as morals, authors' messages, and relations to the readers' lives proved the most difficult for students (McConaughy, 1982).
Structure	1. Clear consistent narrative structure, single point of view, events in chronological order 2. One factor varies (structure, point of view, chronology) 3. Two or more factors vary • Avoid requiring graphics for comprehension for accessibility reasons	1. a. A consistent, linear narrative is maintained throughout the passage. b. The narrative is presented from a single point of view and events are presented in chronological order. 2. a. The passage maintains a clear and focused structure, but with at least one complex element, such as shifts in time, sequence or point of view. b. Changes in structure, point of view, or sequence are well-marked. 3. a. The passage contains multiple elements of	Research shows that texts structured in a linear and/or hierarchical manner are easier to comprehend (Calisir & Gurel, 2003). There are a number of aspects of text structure that affect the ease of comprehension, including shifts in perspective (Fisher, Frey, & Lapp, 2012) and character shifts (Rich & Taylor, 2000). Flashbacks and narrator changes in a story significantly impact readers' abilities to recall or

Difficulty Factor	Passage Metric Description	Grade Level Detail (Sample for Grade 6)	Research-Based Evidence
		complex structure, such as shifts in time, sequence, or point of view. b. Changes in structure, point of view, or sequence are well-marked. c. Elements of structure may contribute to the development of theme, setting, or plot.	retell stories, with more flashbacks and more narrator changes throughout a story compounding this effect (Kucer, 2010).
Language	1. Simple, common word choice, explicit and literal use 2. May include unfamiliar vocabulary, abstract meaning, figurative, ironic, or sarcastic use 3. Generally dense using figurative or purposefully ambiguous, often unfamiliar language	1. a. The passage uses literal, clear, and contemporary language. b. High-frequency, grade-appropriate vocabulary and common word meanings are used. c. Syntax is simple and consistent throughout the passage. d. Interpretation of these words and phrases leads to a singular understanding of their role and meaning within the passage. 2. a. The passage includes some unfamiliar or above-grade-level words. b. The meaning of most or all unfamiliar words can be determined on the basis of context clues. c. Familiar vocabulary may be used to convey figurative meaning. 3. a. The passage includes low-frequency, domain-specific vocabulary, or uncommon word meanings. b. Some variation in syntax may be present. c. The use of figurative, ambiguous, ironic, archaic, or otherwise unfamiliar language to convey meaning is incorporated at this level.	Texts that use common, high-frequency words are easier to understand than texts that use archaic or unfamiliar words. As the amount of familiar vocabulary increases, so does the level of text comprehension (Schmitt, Jiang, & Grabe, 2011). Texts that use unfamiliar language (e.g., Old English) and/or unfamiliar cultural references are more difficult to understand (Fisher, Frey, & Lapp, 2012). Archaic, formal, and domain-specific vocabulary is more difficult than casual or familiar vocabulary (Fisher, Frey, & Lapp, 2012). Both commonness of words and a reader's prior experience impact comprehension. That is, those who read texts with easy vocabulary and are familiar with the topic are able to more easily recall and summarize a text (Freebody & Anderson, 1983). The use of figurative language and meanings also increases the

Difficulty Factor	Passage Metric Description	Grade Level Detail (Sample for Grade 6)	Research-Based Evidence
			difficulty of a text. (Rommers, Dijkstra, & Bastiaansen, 2013). It is easier to understand texts when their words stand for literal meanings. Satire, irony, and allusions are more difficult to interpret than figurative language like imagery or metaphors (Fisher, Frey, & Lapp, 2012).
Total Score	1. Scores below 6 represent low complexity texts. 2. Scores from 6 to 8 represent moderate complexity texts. 3. Scores from 9 to 12 represent high complexity texts.		

The specifications help test developers create or select passages that will support a range of difficulty, furthering the goal of measuring the full range of performance found in the population, but remaining on grade level.

2.2.2 Item Specifications

ELA and mathematics item specifications guide the ICCR item development process. To support the claims in mathematics, the specifications begin by grouping the practices defined in the standards into three practice clusters as follows:

- <u>Practice Cluster 1: Use Mathematics to Solve Problems</u>
  - MP1 – Make sense of problems and persevere in solving them.
  - MP4 – Model with mathematics.
  - MP5 – Use appropriate tools strategically.
- <u>Practice Cluster 2: Use Mathematical Reasoning</u>
  - MP2 – Reason abstractly and quantitatively.
  - MP3 – Construct viable arguments and critique the reasoning of others.
  - MP6 – Attend to precision.
- <u>Practice Cluster 3: Use Characteristics of Problems to Generalize</u>
  - MP7 – Look for and make use of structure.
  - MP8 – Look for and express regularity in repeated reasoning.

Item specifications indicate the mathematics practices implied in each standard. Specifications in mathematics include the following:

- *Content Limits*. This section delineates the specific content measured by the standard and the extent to which the content is different across grade levels. In mathematics, for example, content limits can include acceptable denominators,

number of place values for rounding or computation, acceptable shapes for geometry standards, etc.

- *Acceptable Response Mechanisms*. This section identifies the various ways in which students may respond to a prompt, such as multiple choice, graphic response, proposition response, equation response, and multiple select items. The identified acceptable response mechanisms were identified with accessibility concerns taken into consideration. For example, a graphic response item should only be used when the standard or task demand requires a graphic representation (e.g., graphing a system of equations). Other items, such as multiple choice, can still be used with static images that can be used for all student populations.
- *Mathematics Practice Cluster*. For mathematics, the practices described in the standards have been grouped into clusters of practices. The item specifications outline to which practice cluster (PC) or clusters a particular standard could be aligned: PC1, PC2, PC3, or none.
- *Depth of Knowledge (DOK)*. The task demands of each standard can be classified as DOK 1, DOK 2, or DOK 3. DOK 4 is also used for writing items in ELA. It is important to note that in his recommendations on the assessment of DOK levels for math, Webb did not recommend that DOK 4 items be included in on-demand, state level assessments due to the extended time periods necessary for evaluation. It is recommended that DOK 4 be assessed at the local level (Petit & Hess, 2008).
- *Task Demands*. In this section, the standards are broken down into specific task demands aligned to each standard. Task demands denote the specific ways in which students will provide evidence of their understanding of the concept or skill. In addition, each task demand is assigned appropriate response mechanisms, DOK, and PCs specifically relevant to that particular task demand.
- *Relationship to Range ALDs.* In this section, each task demand is further discussed in light of the Range ALDs. Each task demand corresponds to part of a particular standard, and the discussion of the Range ALDs demonstrates how that task demand relates to a student's level of proficiency with respect to the particular standard (Egan, Schneider, Ferrara, 2012). Range ALDs are defined in Volume 3, Section 4.4.3.
- *Examples and Sample Items*. In this section, sample items are delineated along with their corresponding expected difficulties (easy, medium, and difficult). Notes for modifying the difficulty of each task demand are detailed with suggestions for the item writer. The suggestions for adapting the difficulty based on the task demands are research based and have been reviewed by both content experts and a cognitive psychologist.

Exhibit D presents a sample from the mathematics specifications for one grade 4 standard. Notice that the specification provides guidance for developing items at each acceptable level of DOK, and it identifies the task demands, item types (see Table 1 and Table 2 for brief descriptions of item types, and Appendix E for further detail), and reflection of the

achievement level descriptors to be included at each level. Also note that at each DOK level, the specification provides guidance for developing items in different difficulty ranges.

Exhibit D: Sample Mathematics Specifications for Grade 4

Content Standard	***CCSS.Math.Content.4.NF*** *Number and operations — Fractions* ***Math.Content.4.MD.A*** *Extend understanding of fraction equivalence and ordering* Math.Content.4.NF.A.2 Compare two fractions with different numerators and different denominators (e.g., by creating common denominators or numerators, or by comparing to a benchmark fraction such as ½). Recognize that comparisons are valid only when the two fractions refer to the same whole. Record the results of comparisons with symbols >, =, or <, and justify the conclusions (e.g., using a visual fraction model).
Content Limits	*Denominators are limited to 2, 3, 4, 5, 6, 8, 10, 12, 100. *Benchmarks are limited to 0, ¼, ½, ¾, 1. *Fractions *a*/*b* can be improper fractions and students should not be guided to put fractions in lowest terms or to simplify. *Two fractions being compared should have both different numerator and different denominator.
Calculator	None
Acceptable Response Mechanisms	Equation Response Graphic Response: drag-and-drop (DND), hot spot (HS), drawing Multiple-Choice Response Multiple Select Response Matching Response Editing Task Inline Response Hot Text Draggable Response
Mathematics Practice Cluster	PC1, PC2, PC3
DOK	2, 3
	Model Task
Context	Allowable. Most items at this standard should not have real-world contexts. Any situation that compares two fractions with different numerators and denominators by creating common denominators or numerators or by comparing to benchmark fractions.

DOK Demands

DOK	Task Demand	Response Mechanism	Relationship to Range ALDs	PC1	PC2	PC3	None
DOK 2	1. Compare fractions relating them to benchmark fractions using visual models (e.g., number lines) and/or numeric reasoning.	• Equation response • Graphic response • Multiple-choice response • Multiple select response	Students who can only compare fractions by using benchmark fractions are below or approaching proficient. Similarly, if a student can only compare fractions using visual models, he or she is below or approaching proficient.	x		x	
	2. Interpret information about fractions with different denominators and different numerators to compare fractions using visual models or numeric reasoning.	• Multiple-choice response • Multiple select response	Students who can interpret information about fractions (e.g., their relative sizes) are at or above the proficient level, meaning they have met the standard.	x	x	x	

	3. Compare fractions using symbols <, >, and = with no situational context or visual model.	• Multiple select response • Matching response • Editing task inline response	Students who can fluently compare a variety of fractions using symbols are at the proficient level, meaning they have met the standard.	x		x
	4. Order three or more fractions from least to greatest or greatest to least.	• Hot text draggable response	Students who can extend their fraction comparison thinking by ordering fractions demonstrate an above-proficient level of understanding.			
DOK 3	5. Develop logical arguments, draw conclusions, and relate use of models to numeric strategies to compare fractional quantities	• Equation response • Graphic response • Multiple-choice response • Multiple select response	Depending on the arguments used, a student who performs this task demand could be at varying levels of proficiency. For example, if the logical arguments rely solely on benchmark fractions, then a student is operating at a below or approaching proficient achievement level. Conversely, if a student is fluently comparing fractions and flexibly working with various types of models and fractions (e.g., improper fractions) then the student is operating at a proficient or highly proficient level.	x	x	x

	Example
Context	Compare fractions, or fractions represented by models, with or without a situational context, such as pizza. • A fraction's denominator does not have to be a multiple of the other (e.g., 2/5 and 2/3). • Fractions are less than 1. • Both fractions can be non-unit fractions.
Context easier	• Fractions are less than 1. • One of the fractions involved is a unit fraction. • One fraction's denominator is a multiple of the other.
Context more difficult	• One or both are improper fractions.

Item Models	Sample Item	Difficulty	PC	Response Mechanism	Notes, Comments
DOK 2	Select >, <, or = to complete a true statement about each pair of fractions. 1/2 ☐ 3/8 [include at least two more pairs of fractions]	Easy	1, 2	Matching response	This is a DOK 2 because students are comparing fractions using <, >, or =. It is easy because both fractions are less than 1, and one fraction is a unit fraction.
	Select >, <, or = to complete a true statement about each pair of fractions. 3/5 ☐ 5/12 [include at least two more pairs of fractions]	Medium	1, 2	Matching response	This is a DOK 2 because students are comparing fractions using <, >, or =. It is medium because both fractions are less than 1.
	Select >, <, or = to complete a true statement about each pair of fractions. 4/3 ☐ 6/5	Hard	1, 2	Matching response	This is a DOK 2 because students are comparing fractions using <, >, or =.

	[include at least two more pairs of fractions]				It is hard because both fractions are “improper” fractions.
DOK 3	Kari has two fraction models, each divided into equal-sized sections. The fraction represented by Model Q is greater than the fraction represented by Model R. Part A. Generate Model Q so it is divided into 8 sections, and 5 sections are shaded. Then, generate Model R so it is divided into 12 sections. Part B. Complete the fraction comparison statement. Part C. Which statement is true about the two fraction models you generated and the comparison between them?	Medium	1, 2, 3	• Simulation response • Editing task inline response • Multiple-choice response	This is a DOK 3 because students have to develop logical arguments, draw conclusions from given information, and relate use of models to numeric strategies to compare fractional quantities. It is medium because students have to construct models using same-sized wholes and then complete a true comparison between the fractional quantities. Both fractions are not unit fractions.

Similar to mathematics, the ELA item specifications include the following information:

- *Content Standard.* This identifies the standard being assessed.
- *Content Limits.* This section delineates the specific content that the standard measures and the parameters in which items must be developed to assess the standard accurately, including the lower and upper complexity limits of items.
- *Acceptable Response Mechanisms.* This section identifies the various ways in which students may respond to an item or prompt. Here, we note whether evidence-based selected-response (two-part items), extended response, hot text, multiple choice, multiple select, and/or short answer (to be scored automatically with our *proposition scorer* [the machine scoring procedure based on the detailed rubric]) items may be used, and if so, how.
- *DOK Demands.* This section is broken into three subsections: DOK, task demand, and response mechanism. The task demands explain the skills the students may be required to demonstrate and connect these skills to each applicable DOK. The task demands break down the cognitive complexity to show how each DOK level requires differences in higher-order thinking. Finally, the DOK and task demand are connected to appropriate response mechanisms used to assess these skills.

- *Sample Items.* In this section, sample items present a range of response mechanisms and their corresponding expected difficulties (easy, medium, and hard). Notes delineating the cognitive demands of the item and an explanation of its difficulty level are detailed for each sample item.

Exhibit E is a sample of the item specifications our content experts developed for a grade 6 literacy standard. It outlines the limits of the item content to fully address the standard. This includes specifying the type and amount of evidence required. Furthermore, as the standard requires citing "several pieces of textual evidence," we limit the acceptable response mechanisms to hot text, wherein the student selects the evidence in the text itself, and multiple select, which allows students to choose two or more disparate pieces of evidence. The DOK sections explain the demands for each DOK level and provide the acceptable response mechanisms. The cognitive demands increase from supporting an explicit inference with explicit evidence (DOK 1) to providing implicit evidence for an inference that the student makes (DOK 3). This level of detail provides the item writer with guidance when developing items, ensuring that the items address the standard and are correctly aligned at the DOK and difficulty levels. DOK 4 is a special category for only writing prompt items, and as such is not included in these item specifications.

Exhibit E: Sample ELA Item Specification for Grade 6

Content Standard	**Literacy RL.6.1: Cite textual evidence to support analysis of what the text says explicitly as well as inferences drawn from the text.**			
Content Limits	Items may ask for text-based evidence to support what is directly stated in the text. Items may ask the student to find evidence to support an inference made by the item writer or by the student.			
Acceptable Response Mechanism	Hot Text • Requires the student to select words or phrases from the text to answer questions using explicit information in the text as support. • Requires the student to select an inference from four choices and then to select words or phrases from the text to support the inference (two-part Hot Text). Multiple Choice • Requires the student to select from four choices to answer questions using explicit or implicit information from the text as support.			
DOK	1, 2			
	DOK Demands			
DOK	Task demand		Response mechanism	
DOK 1	Identify support for a statement in the text where both the statement and support are explicit.		1. Hot Text Response 2. Multiple Choice Response	
DOK 2	Provide text-based support for an inference drawn from the text. The item writer may or may not provide the inference for the student.		1. Hot Text Response 2. Multiple Choice Response	
DOK 3	N/A			
Item Models	Sample Item	Difficulty	Notes, Comments	Passage

DOK 1	Select the sentence from the paragraph that shows why Papa had to leave the farm to go work on the railroad. [Hot Text]	Easy	The student must understand that the price of cotton dropped, meaning the family did not have enough money. The text explicitly states the answer to the question and the student does not need to wade through extraneous details. The item difficulty is easy because the support directly precedes the idea in the text. Easy Difficulty: The answer is explicitly stated in the text.	*Roll of Thunder, Hear My Cry*
DOK 1	Where does Brian get the idea about how to store live fish in the water? [Multiple Choice]	Medium	The student must identify which detail in the text gives Brian the idea of how to store the fish. Although the answer is stated explicitly in the text, the student must sort through multiple details and paragraphs, increasing the difficulty of the item. The student must make a connection between the woven door Brian uses for his food shelter and the gate he uses to close off part of the river, trapping the fish inside. Medium Difficulty: The answer is explicitly stated, but the information must be combined from details in several paragraphs.	*Hatchet*
DOK 2	Which sentence from the text shows that the family's financial situation has not improved? [Multiple Choice]	Easy	The student must use details from the text to show that the family's financial situation still has not improved. The item difficulty is easy because the inference is provided for the student and the support is directly stated in the text. The student must choose the correct support from four answer choices. Easy Difficulty: The support for the inference stated in the question is explicitly provided in the text.	*Roll of Thunder, Hear My Cry*

DOK 2	Select a sentence from the text that shows that the family's financial situation has still not improved. [Hot Text]	Medium	The student must support an inference provided by the item. The inference that the family's financial situation has not improved is provided. The student must infer that because Papa is returning to work on the railroad again, the family still needs to raise money beyond what they earn from the farm. The student must select an example embedded within the text, increasing the number of options and, thus, the difficulty of the item. Medium Difficulty: The student must choose which sentence (among all the sentences in the text) supports the inference provided in the question.	*Roll of Thunder, Hear My Cry*
DOK 2	Reread paragraph 6. Part A: Why does Papa believe the land is so important? Part B: Select the sentence from the text that shows why Papa thinks the land is so important. [two-part Hot Text]	Hard	The item requires the student to interpret details from the text to recognize Papa's reason for believing the land is so important. The student must differentiate between the description of the land, Cassie's thoughts and feelings, and quotes from Papa. In Part B, the student must integrate details from across the text to draw an inference about the importance of the land. The student must recognize that owning the land means that the family does not have to answer to anyone else. This item is difficult because the student must draw inferences and interpret multiple details from the text. Hard Difficulty: The student must infer the answer to the question based on a character's dialogue and then select a sentence from the text that supports this inference.	*Roll of Thunder, Hear My Cry*

2.3 Selection and Training of Item Writers

All item writers developing ICCR items have at least a bachelor's degree, and many bring teaching experience. All item writers are trained in

- the principles of universal design;
- the appropriate use of item types; and
- the ICCR specifications.

Key materials are included in Appendix G. These include

- CAI's Language Accessibility, Bias, and Sensitivity (LABS) Guidelines; and
- a training (presented using Microsoft PowerPoint) for the appropriate use of item types.

Sample specifications for passages, mathematics, and ELA are presented in Exhibits C, D, and E respectively.

2.4 Internal Review

ICCR's test development structure utilizes highly effective units organized around each content area. Unit directors oversee team leaders who work with team members to ensure item quality and adherence to best practices. All team members, including item writers, are content-area experts. Teams include senior content specialists who review items prior to client review and provide training and feedback for all content-area team members.

ICCR items undergo a rigorous, multiple-level internal review process before they are sent to external review. Staff members are trained to review items for both content and accessibility throughout the entire process. A sample of the item review checklist that our test developers use is included in Appendix F. The ICCR internal review cycle includes the following phases:

- Preliminary Review
- Content Review One
- Edit Review
- Senior Review

2.4.1 Preliminary Review

Preliminary review is conducted by team leads or senior content staff. Sometimes preliminary review is conducted in a group setting, led by a senior test developer. During the preliminary review process, test developers, either individually or as a group, analyze items to ensure the following:

- The item aligns with the academic standard.
- The item matches the item specification for the skill being assessed.

- The item is based on a quality idea (i.e., it assesses something worthwhile in a reasonable way).
- The item is properly aligned to a DOK level.
- The vocabulary used in the item is appropriate for the grade and subject matter.
- The item considers language accessibility, bias, and sensitivity.
- The content is accurate and straightforward.
- The graphic and stimulus materials are necessary to answer the question.
- The stimulus is clear, concise, and succinct (i.e., it contains enough information to know what is being asked, it is stated positively, and it does not rely on negatives—such as *no, not, none, never*—unless absolutely necessary).

For selected-response items, test developers also check to ensure that the set of response options are

- as succinct and short as possible (without repeating text);
- parallel in structure, grammar, length, and content;
- sufficiently distinct from one another;
- all plausible (but with only correct option); and
- free of obvious or subtle cueing.

For machine-scored constructed-response items, item developers also check that the items score as intended at each score point in the rubric and that scoring assertions address the skill that the student is demonstrating with each type of response.

At the conclusion of the Preliminary Review, items that were accepted as written or revised during this review move on to Content Review One. Items that were rejected during this review do not move on.

2.4.2 Content Review One

Content Review One is conducted by a senior content specialist who was not part of the Preliminary Review. This reviewer carefully examines each item based on all the criteria identified for Preliminary Review. He or she also ensures that the revisions made during the Preliminary Review did not introduce errors or content inaccuracies. This reviewer approaches the item both from the perspective of potential clients and from his or her own experience in test development.

2.4.3 Edit Review

During the Edit Review, editors have four primary tasks.

First, editors perform basic line editing for correct spelling, punctuation, grammar, and mathematical and scientific notation, ensuring consistency of style across the items.

Second, editors ensure that all items are accurate in content. Editors compare reading passages against the original publications to make sure that all information is internally consistent across stimulus materials and items, including names, facts, or cited lines of text that appear in the item. Editors ensure that the keys are correct and that all information in the item is correct. For mathematics items, editors perform all calculations to ensure accuracy.

Third, editors review all material for fairness and language accessibility issues, using CAI's Language Accessibility, Bias, and Sensitivity (LABS) Guidelines shown in Appendix G.

Finally, editors confirm that items reflect the accepted guidelines for good item construction. In all items, they look for language that is simple, direct, and free of ambiguity with minimal verbal difficulty. Editors confirm that a problem or task and its stem are clearly defined and concisely worded with no unnecessary information. For multiple-choice items, editors check that options are parallel in structure and fit logically and grammatically with the stem and that the key accurately and correctly answers the question as posed, is not inappropriately obvious, and is the only correct answer to an item among the distractors. For constructed-response items, editors review the rubrics for appropriate style and grammar.

2.4.4 Senior Review

By the time an ICCR item arrives at Senior Review, it has been thoroughly vetted by both content reviewers and editors. Senior reviewers (in particular, Senior Content Specialists) look back at the item's entire review history, making sure that all the issues identified in that item have been adequately addressed. Senior reviewers verify the overall content of each item, confirming its accuracy, alignment to the standard, and consistency with the expectations for the highest quality. For machine-scored constructed-response items, senior reviewers carefully check the rubric and scoring logic by responding to the task just as the student would in the testing environment. They check full-credit, partial-credit, and zero-credit responses to verify that the scoring is working as intended and that the scoring assertions adequately address the evidence the student provides with each type of response.

2.5 Review by State Personnel and Stakeholder Committees

All ICCR items have been through an exhaustive external review process. Items in the bank were reviewed by content experts in several states and reviewed and approved by multiple stakeholder committees to evaluate both content and bias/sensitivity.

2.5.1 State Review

After items have been developed in the ICCR item bank, state content experts review any eligible items prior to committee review. At this stage in the review process, clients can request edits, such as wording edits, scoring edits, or alignment or DOK updates. A CAI Director for Mathematics or ELA reviews all client-requested edits in light of the ICCR item specifications, other clients' requests, and existing items in the bank to determine whether or

not the requested edits will be made. At this stage, clients have the option to present these items to committee (based on the edits made) or withhold them from committee review.

For items that have already been field tested in other states, wording and scoring edits cannot be made (as such edits risk altering the function of calibrated items), and clients can simply select the items from the available item bank to present to committee.

2.5.2 Content Advisory Committee Review

During the Content Advisory Committee Reviews, items are reviewed for content validity, grade-level appropriateness, and alignment to the content standards. Content Advisory Committee members are typically grade-level and subject-matter experts, or they may include mathematics coaches (who can speak to standards across grades) or literacy specialists. During this review, educators also ensure that the rubrics for machine-scored constructed-response items reflect the anticipated correct responses (see more information in the Rubric Validation section).

A summary of the committee meetings appears in Exhibit F, with further details about the participants and information regarding later meetings, in Appendix H.

Exhibit F: Summary of ICCR Content Advisory Committee Meetings

Location	Year	Number of Committee Members	Number of Items Reviewed
Arizona	2014	78	2,850
	2015	52	871
	2016	40	1,072
	2017	43	918
	2018	36	911
Utah	2014	56	1,139
	2015	53	879
	2016	60	352
	2017	36	506
Florida	2014	108	1,765
	2015	122	963
	2016	56	524
	2017	78	528
New Hampshire	2018	29	257
North Dakota	2018	30	319
West Virginia	2018	24	317
Wyoming	2018	36	503

2.5.3 Language Accessibility, Bias, and Sensitivity Committee Review

During the bias and sensitivity reviews, stakeholders review items to check for issues that might unfairly impact students based on their background. For example, some states include representatives from the special education, low vision, hearing impaired, and other student populations. Further, diverse members of this committee represent students of various ethnic and economic backgrounds to ensure that all items are free of bias and sensitivity concerns.

A summary of the committee meetings appears in Exhibit G, with additional details about the participants and information regarding later meetings in Appendix I.

Exhibit G: Summary of Fairness Committee Meetings

Location	Year	Number of Committee Members	Number of Items Reviewed	Number of Items Rejected
Florida	2015	32	1,147	0
	2016	22	1,065	9
	2017	28	392	0
Utah	2015	21	2,626	96
	2016	65	595	11
	2017	13	575	13
Arizona	2015	25	786	1
	2016	20	1,113	15
	2017	20	926	0
	2018	20	899	1
New Hampshire	2018	30	261	0
North Dakota	2018	8	340	10
West Virginia	2018	15	853	1
Wyoming	2018	36	507	0

2.5.4 Markup for Translation and Accessibility Features

After all approved state and committee recommended edits have been applied, the items are considered “locked” and ready for all accessibility tagging. Accessibility markup is embedded into each item as part of the item development process rather than as a post hoc process applied to completed tests.

Accessibility markups, whether they are translations or they are used for text-to-speech, follow similar processes. One trained expert enters the markup. A second expert reviews the work and recommends changes if necessary. If there is disagreement, a third expert is engaged to resolve the conflict.

Currently, ICCR items are tagged with Spanish translations and text-to-speech, including Spanish text-to-speech.

2.6 FIELD TESTING

ICCR items were field tested by being embedded in operational, summative, and accountability assessments in participating states. The field testing is described in detail in Volume 1, Section 3.2.

2.7 POST FIELD-TEST REVIEW

Following field testing, items were subject to additional reviews. These included

- key verification, for items that are key-scored;
- rubric validation, for machine-scored items that are rule based or heuristic based;
- rangefinding for essays; and
- data review for items that failed standard flagging criteria.

We discuss each of these processes in the following paragraphs.

2.7.1 Key Verification

Key verification is a simple process by which we create a frequency table of response frequencies and the scores that they received. These are reviewed by qualified content staff to ensure that all correct responses, and only correct responses, receive a score.

2.7.2 Rubric Validation

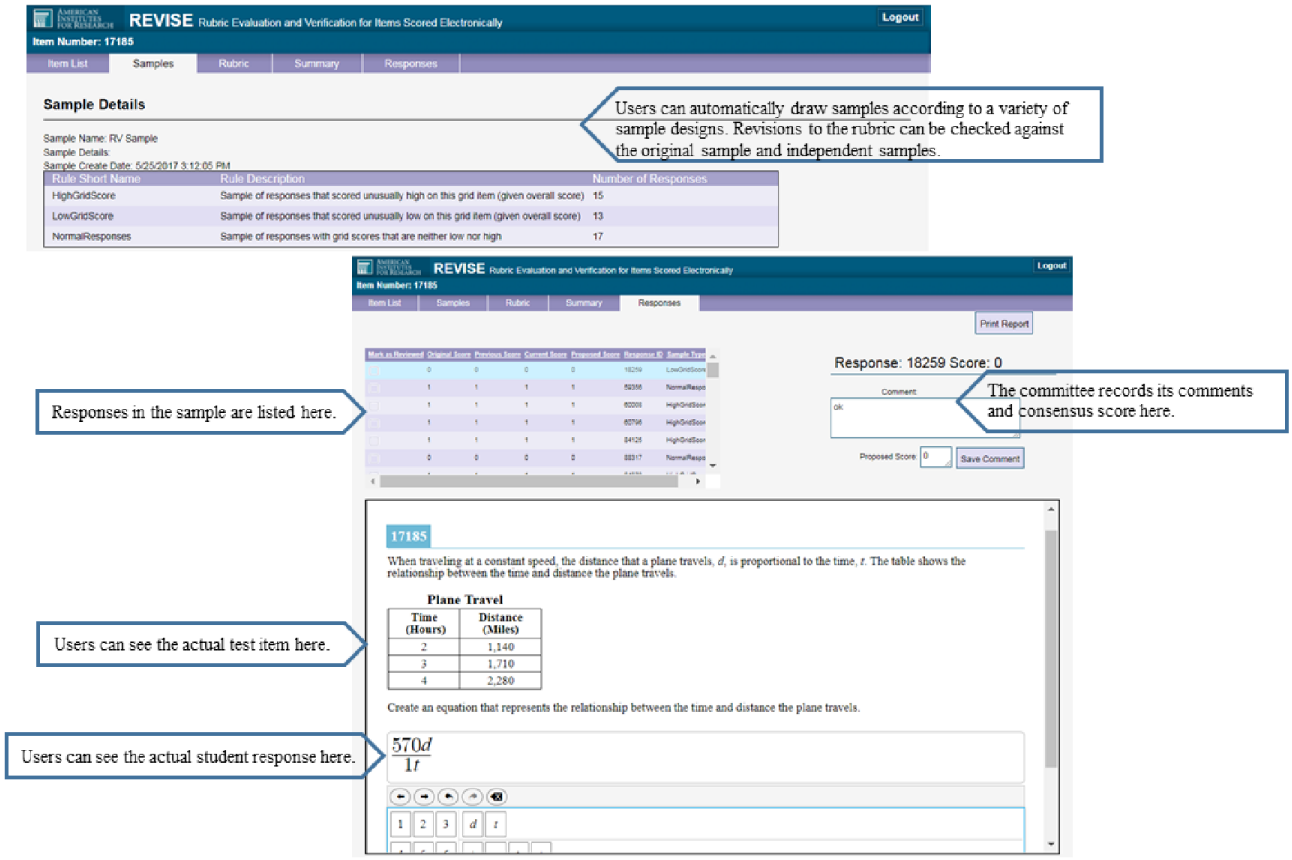
More complex selected-response items, as well as machine-scored constructed-response items, undergo rubric validation, which occurs in two phases. During the first phase, CAI content experts draw one or more samples to identify anomalous or unforeseen responses and ensure that they are scored correctly. The rubrics may be adjusted, and responses rescored, at this point.

The second phase of rubric validation involves ICCR or state content experts. During this phase, a fresh sample of responses is drawn from three strata in equal numbers: low-scoring responses from otherwise high-scoring students, high-scoring responses from otherwise low-scoring students, and a random sample from the remainder.

During these reviews, experts review responses and scores in a system we call *REVISE*. Items are reviewed as the students saw them, along with the students' responses. The experts' comments are captured, and rubrics are accepted or updated as consensus is reached. Often, these discussions adjust tolerances. For example, in drawing a best-fitting line, the experts may choose to be more or less lenient in accepting a line as "close enough." In this regard, the process is similar to rangefinding.

Exhibit H shows some features from REVISE.

Exhibit H: Features of the REVISE Software



ITS archives critical information regarding the scoring certification completed during the rubric validation process. This includes any rubric changes made during the scoring decision meetings and the sign-off completed by the senior content expert once the rubric has been changed, rescoring has been completed, and it has been verified that the scoring using the final rubric functioned as intended.

Following rubric validation, all items are subject to statistical checks, and flagged items are presented in data review committees.

2.7.3 Rangefinding

Items requiring hand scoring undergo a committee process called *rangefinding,* which engages educators and content experts in interpreting the rubric and selecting exemplars that will be used to train and validate hand scoring. Hand scoring results were used to train scoring engines. This process is discussed in Volume 4, along with the details of the rangefinding efforts.

2.7.4 Data Review

Volume 4, Section 6.1 describes in detail the statistical flags that send items to data review. The flags are designed to highlight potential content weaknesses, miskeys, or possible bias issues. Committee members are taught to interpret these flags and given guidelines for examining the

items for content or fairness issues. A sample of the training materials used for these data review meetings appears in Appendix J.

Exhibit I summarizes the data review committee meetings. Details, including the composition of each committee, appear in Appendix K.

Exhibit I: Summary of Data Review Committee Meetings

Location	Year	Number of Committee Members	Number of Items Reviewed	Number of Items Rejected
Utah	2015	60	1,139	0
	2016	82	879	17
	2017	68	352	22
Arizona	2017	43	1,072	25
	2018	40	918	38

3. ICCR ITEM BANK SUMMARY

The ICCR item bank is robust and has been constructed explicitly to support multiple statewide assessment programs. As described above, ICCR items were written to the College- and Career-Ready Standards, and the bank is occasionally augmented with items measuring some state-specific standards. The ICCR item bank is designed to be sufficiently robust to support a range of test designs, including item-adaptive, multi-stage adaptive, and fixed-form tests.

Each state using the ICCR item bank selects items for use on its statewide assessment from those that are appropriately aligned and have passed required reviews (as described in Section 2). The ICCR continues to grow as CAI continues to field test new items in participating states. Participating states collectively share the items and agree to field test new items each year. Summaries of current item inventories are provided below.

3.1 CURRENT COMPOSITION OF THE ITEM BANK

Table 1 and Table 2 list the ELA and mathematics item types and provide a brief description of each item type. Additional information about the item types can be found in Appendix A for ELA and Appendix B for mathematics. Examples of various item types can be found in Appendix E.

Table 1: 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.
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.

Response Type*	Description
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.
Table 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.

**Note that the abbreviations correlate to the attributes used in* CAI*'s Item Tracking System.*

Table 2: 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.
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.
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 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.
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.

**Note that the abbreviations correlate to the attributes used in CAI's Item Tracking System.*

Table 3 through Table 15 provide the number of items and writing prompts in the ICCR item bank available for use in spring 2021 statewide assessments.

Table 3: ICCR ELA Spring 2021 Operational and Field-Test Item Pool

Grade	Total Number of Items	Number of Writing Prompts
3	520	6
4	552	6
5	535	6
6	609	7
7	611	6
8	596	6
9	370	5
10	355	3

Table 4: ICCR ELA Spring 2021 Operational Item Pool

Grade	Number of Total OP Items
3	431
4	482
5	453
6	558
7	546
8	509
9	330
10	308
Total	**3,617**

Table 5: ICCR ELA Spring 2021 Field-Test Item Pool

Grade	Number of Total FT Items
3	89
4	70

Grade	Number of Total FT Items
5	82
6	51
7	65
8	87
9	40
10	47
Total	**531**

Table 6: ICCR ELA Spring 2021 Item Counts by Grade and Reporting Category

Grade	Reading Informational Text	Reading Literary Text	Writing and Language	Speaking and Listening	Grand Total
3	214	173	112	7	506
4	204	185	143	7	539
5	196	175	139	11	521
6	258	188	134	15	595
7	236	213	139	10	598
8	247	189	142	5	583
9	146	135	81	8	370
10	155	105	92	2	354
Total	**1,656**	**1,363**	**982**	**65**	**4,066**

Table 7: ICCR ELA Spring 2021 Item Counts by Grade and Depth of Knowledge

Grade	DOK 1	DOK 2	DOK 3	DOK 4	Grand Total
3	91	337	72	6	506
4	95	366	73	6	540
5	90	352	74	6	522
6	89	393	106	7	595
7	82	398	112	6	598
8	83	388	106	6	583
9	38	259	68	5	370
10	53	229	69	3	354
Total	**621**	**2,722**	**680**	**45**	**4,068**

Table 8: ICCR ELA Spring 2021 Item Counts by Grade and Item Type

Grade	Item Type	Number of Items
3	Editing Task Choice	53
	Hot Text	41
	Multiple Choice	383
	Multiple Select	32
	Table Match	15
	Text Entry	10
	Total	***534***
4	Editing Task Choice	69
	Hot Text	50
	Multiple Choice	386
	Multiple Select	47
	Table Match	13
	Text Entry	14
	Total	***579***
5	Editing Task Choice	78
	Grid	1
	Hot Text	61
	Multiple Choice	345
	Multiple Select	50
	Table Match	19
	Text Entry	10
	Total	***564***
6	Editing Task Choice	74
	Hot Text	49
	Multiple Choice	439
	Multiple Select	55
	Table Match	12
	Text Entry	12
	Total	***641***
7	Editing Task Choice	73
	Hot Text	46
	Multiple Choice	415
	Multiple Select	104
	Table Match	7
	Text Entry	12
	Total	**657**

Grade	Item Type	Number of Items
8	Editing Task Choice	71
	Hot Text	47
	Multiple Choice	433
	Multiple Select	64
	Table Match	9
	Text Entry	8
	Total	***632***
9	Editing Task Choice	37
	Grid	1
	Hot Text	41
	Multiple Choice	259
	Multiple Select	29
	Table Match	2
	Text Entry	7
	Total	***376***
10	Editing Task Choice	47
	Hot Text	36
	Multiple Choice	249
	Multiple Select	23
	Text Entry	4
	Total	**359**
All	***Grand Total***	**4,342**

Table 9: ICCR Mathematics Spring 2021 Operational and Field-Test Item Pool

Grade	Total Number of Items
3	661
4	696
5	571
6	709
7	493
8	574
HS	1,344
Total	**5,048**

Table 10: ICCR Mathematics Spring 2021 Operational Item Pool

Grade	Number of Sp21 OP Items
3	613
4	649
5	526
6	658
7	453
8	524
HS	1,296
Total	**4,719**

Table 11: ICCR Mathematics Spring 2021 Field-Test Item Pool

Grade	Number of Sp21 FT Items
3	48
4	47
5	45
6	51
7	40
8	50
HS	49
Total	**330**

Table 12: ICCR Mathematics Spring 2021 Item Counts by Grade and Reporting Category

Grade	Reporting Category	Number of Items
3	Geometry	51
	Measurement and Data	135
	Number and Operations—Fractions	159
	Number and Operations in Base Ten	109
	Operations and Algebraic Thinking	187
	Total	***641***
4	Geometry	62
	Measurement and Data	107

Grade	Reporting Category	Number of Items
	Number and Operations—Fractions	200
	Number and Operations in Base Ten	186
	Operations and Algebraic Thinking	121
	Total	***676***
5	Geometry	67
	Measurement and Data	74
	Number and Operations—Fractions	171
	Number and Operations in Base Ten	148
	Operations and Algebraic Thinking	92
	Total	***552***
6	Expressions and Equations	207
	Geometry	78
	Ratios and Proportional Relationships	165
	Statistics and Probability	66
	The Number System	173
	Total	***689***
7	Expressions and Equations	88
	Geometry	97
	Ratios and Proportional Relationships	92
	Statistics and Probability	108
	The Number System	88
	Total	***473***
8	Expressions and Equations	161
	Functions	113
	Geometry	140
	Statistics and Probability	77
	The Number System	63
	Total	***554***
HS	Algebra	328
	Functions	357
	Geometry	410
	Number and Quantity	81
	Statistics and Probability	166
	Total	**1,342**
All	***Grand Total***	**4,927**

Table 13: ICCR Mathematics Spring 2021 Item Counts by Grade and Depth of Knowledge

Grade	DOK 1	DOK 2	DOK 3	Total
3	150	405	86	641
4	156	440	80	676
5	112	364	76	552
6	177	445	67	689
7	85	308	80	473
8	127	324	103	554
HS	178	990	174	1,342
Total	**985**	**3,276**	**666**	**4,927**

Table 14: ICCR Mathematics Spring 2021 Item Counts by Item Type

Grade	Item Type	Number of Items
3	Editing Task Choice	11
	Equation	363
	Grid	92
	Multiple Choice	138
	Multiple Select	57
	Table Input	15
	Table Match	13
	Text Entry	1
	Total	***690***
4	Editing Task Choice	19
	Equation	380
	Grid	61
	Multiple Choice	114
	Multiple Select	101
	Table Input	16
	Table Match	33
	Total	***724***
5	Editing Task Choice	19
	Equation	351
	Grid	35
	Multiple Choice	113
	Multiple Select	51

Grade	Item Type	Number of Items
	Table Input	11
	Table Match	15
	Text Entry	1
	Total	***596***
6	Editing Task Choice	16
	Equation	369
	Grid	49
	Multiple Choice	196
	Multiple Select	59
	Table Input	30
	Table Match	17
	Total	***736***
7	Editing Task Choice	12
	Equation	312
	Grid	39
	Multiple Choice	120
	Multiple Select	21
	Table Input	3
	Table Match	10
	Total	***517***
8	Editing Task Choice	19
	Equation	266
	Grid	60
	Multiple Choice	187
	Multiple Select	56
	Table Input	8
	Table Match	10
	Total	***606***
HS	Editing Task Choice	25
	Equation	590
	Grid	65
	Multiple Choice	542
	Multiple Select	74
	Table Input	8
	Table Match	16
	Total	**1,320**
All	***Grand Total***	**5,189**

Table 15: Spring 2021 ICCR Cluster Item Counts

Grade	Clusters
ELA	
3	14
4	12
5	13
6	14
7	13
8	13
HS	1
Mathematics	
3	20
4	20
5	19
6	20
7	20
8	20
HS	2

3.2 Strategy for Pool Evaluation and Replenishment

CAI seeks to release approximately 5% of the pool each year, although the actual number of items released depends on client needs in any given year. CAI intends to field test an additional 10–15% of the pool each year, seeking to grow the pool over time.

Items are field tested each year in embedded field test slots. Field testing is described in detail in Volume 1, Section 3.2. Currently, writing prompts are field tested in independent field tests approximately every five years.

Our general strategy for targeting item development gathers information from three sources:

1. Tabulations of content coverage and ranges of difficulty to identify gaps in the pool
2. Characteristics of released items to be replaced
3. Characteristics of items overused in adaptive programs

Each year, before an adaptive test goes live, simulations are used to fine-tune the parameters of the adaptive algorithm. This fine-tuning optimizes the balance between blueprint match and individualized information. Among the many reports from the simulator are items that are seen by more than 20% of students. The characteristics of these items are the primary targets for development. Overused items become candidates for release two years hence, once replacements have been introduced into the operational pool.

4. North Dakota State Assessment Test Construction

Using ICCR as the source of items for the NDSA in ELA and mathematics, tests in North Dakota were constructed to meet the state-specific test blueprints that were written to align with the North Dakota Content Standards. Because the ICCR item bank is large and contains an array of item types, the tests could be uniquely developed by drawing from the pool of available ICCR items. The construction of test item pools is a process that requires both expert judgment from content experts and examination of psychometric criteria to ensure that certain technical characteristics of the tests meet industry expected standards. The processes used for blueprint development and test item pool construction are described to support the claim that they are technically sound and consistent with expectations of current professional standards.

The NDSA is designed to support the claims described in this volume.

4.1 Test Blueprints

CAI worked closely with the North Dakota Department of Public Instruction (NDDPI) to create blueprints that guided the development process for the NDSA in ELA and mathematics grades 3−8 and 10. Our blueprints were designed to meet the following objectives:

- Full coverage of the breadth and depth of North Dakota's standards
- Less than five hours of total testing time, including 60 minutes of writing
- All machine-scored items, including many true constructed-response item types, in which students must construct an equation, graph, illustration, etc.

4.1.1 ELA Blueprints

The blueprints developed for English language arts are provided in Appendix A: English Language Arts Blueprints. The blueprints are organized by strand and specify the number of items required for each reporting category, ensuring that the test contains sufficient items at that category to elicit enough information from the student to justify strand-level scores. Appendix A also shows the reporting categories and required number of items in the proposed ELA blueprints.

The ELA blueprint results in a test design that delivers the following to each student:

- Two informational reading passages with associated items
- Two literary reading passages with associated items
- Eight to ten language items
- One text-based writing task

The blueprint defines the reading sub-strands and individual standards within each sub-strand. The blueprint also defines the individual standards within the Language and Writing reporting categories. The sub-strands and standards have assigned item ranges to ensure that the material is represented on a test with the proper emphasis relative to other standards in that reporting category. The item ranges for individual standards ensure that at least half of the standards in any reporting

category or sub-strand must be represented on a test. The item ranges in the blueprint allow each student to experience a wide range of content while still providing flexibility during test construction. Writing is measured by an extended text-based writing task representing the writing dimensions of Organization/Purpose, Evidence/Elaboration, and Conventions. The ELA blueprint also includes ranges for DOK. The blueprints require 16−21% of items on a test to be DOK 1, 18−29% of items to be DOK 3 or higher, and the rest are DOK 2. These ranges are included in Table 26 and Table 27 for ELA and mathematics, respectively.

Because the ICCR item bank offers a range of item types to assess all the standards described here, each test constructed fulfills the NDSA blueprint with a variety of item types that capitalize on efficiency while providing a deep measure of the content standards. The blueprints ensure coverage of the breadth and depth of the standards while reducing testing time.

CAI estimated ELA testing time to be within 90 minutes for students whose testing time falls within the 85th percentile, as shown in Table 16. To estimate these times, CAI analyzed the average testing time for students on the 2015−2016 North Dakota test. CAI computed the average page time per item for reading literary and informational passages and multiplied that average page time per item by the number of informational or literary items specified in the blueprint. These time estimates represent the testing time for two literary passages, two informational passages, their associated items, and language items. The observed testing times in Table 17 represent the 2021 test administration for the NDSA ELA portion. Both reading and writing times are significantly shorter than the estimated times. Observed NDSA testing times will be continually monitored over time with the expectation that observed times may change as familiarity with the NDSA improves.

Table 16: Estimated ELA Testing Times (In Minutes)

Subject	Grade	85th Percentile Testing Time
Reading	3	150
	4	120
	5	130
	6	125
	7	125
	8	105
	10	90
Writing	3	90
	4	90
	5	90
	6	90
	7	90
	8	90
	10	90

Table 17: Observed ELA Testing Times (In Minutes)

Subject	Grade	Mean Testing Time*	85th Percentile Testing Time
Reading	3	67	95
	4	56	79
	5	61	85
	6	57	79
	7	54	75
	8	49	68
	10	44	60
Writing	3	50	78
	4	54	85
	5	53	81
	6	48	77
	7	42	65
	8	39	60
	10	39	60

4.1.2 Mathematics Blueprints

The blueprints developed for mathematics are shown in Appendix B. They are organized by content domain. Reporting categories at a specific grade consist of a single content domain or, when necessary and appropriate, a combination of content domains. For each reporting category, the blueprints specify a minimum and maximum number of items on each test that should contribute to that category. This ensures that the test contains enough items at that category to elicit enough information from the student to generate an ability estimate. Appendix B also shows the reporting categories and item ranges in the proposed mathematics blueprints.

Within a reporting category, the blueprint defines content clusters that contain varying numbers of related content standards. Both the content clusters and underlying content standards are assigned item ranges. The item ranges for the content clusters ensure that that material is represented on a test with the proper emphasis relative to other clusters in that reporting category. The item ranges for individual standards are constructed so that at least one-half of the standards in any particular content cluster must be represented on a test. The item range approach ensures that all tests expose students to a wide range of content in the correct proportion while providing some flexibility during test construction. The North Dakota Mathematics Content Standards do not identify specific standards that should be included in a grade 10 test. On the grade 10 blueprint, for instance, we began with standards and function types (generally linear and exponential) that are appropriate for an Integrated Mathematics course. We then modified that blueprint, including important basic standards involving quadratic functions. The blueprint also includes geometry and statistics and

probability. This results in a blueprint that is strongly rooted in North Dakota's foundational high school standards and appropriate for grade 10. Although the current item bank does allow for blueprint coverage, additional item development is underway to ensure all geometry high school standards are assessed appropriately. The mathematics blueprints also contain item ranges for DOK. In general, the blueprints require 17%−26% of items on a test to be DOK 1, 14%−26% of items to be DOK 3, and the rest to be DOK 2. These item ranges ensure that all students are exposed to varying levels of cognitive complexity while still providing some flexibility during test construction.

The ICCR item bank contains many different item types, such as traditional multiple-choice items, technology-enhanced items, and machine-scored constructed-response items. Any test built from this bank will have a wide variety of item types represented. Thus, we did not place artificial restrictions on the number of each specific item type that a particular test must contain, and the sample blueprints contain no such restrictions.

We estimated that the testing time for mathematics would be well within 90 minutes. To estimate these times, we first looked at the average testing time of students on typical ICCR mathematics items. In general, across all grades, students spent more time on machine-scored constructed-response items than on selected-response items. We then looked at the proportion of each specific item type with regard to the item type category within the ICCR item bank. Using these numbers, we arrived at the average time spent on selected-response and machine-scored constructed-response items, given the composition of the item bank. Based on these averages and the number of items per test, we arrived at the estimated mathematics testing times provided in Table 18. The observed testing times in Table 19 represent the 2021 test administration for the NDSA mathematics portion and are substantially less than originally predicted across all grades. NDSA testing times will be continually monitored over time with the expectation that observed times may change as familiarity with the NDSA improves.

Table 18: Estimated Mathematics Testing Times (In Minutes)

Grade	85th Percentile Testing Time
3	135
4	140
5	155
6	155
7	145
8	135
10	135

Table 19: Observed Mathematics Testing Times (In Minutes)

Grade	Mean Testing Time	85th Percentile Testing Time
3	57	81
4	58	82
5	62	84
6	63	87
7	58	79
8	55	74
10	52	70

4.1.3 NDSA Test Specifications

For each grade level, one ELA and one mathematics item pool was constructed using a pre-equated design. With the pre-equated design, all item parameters from the item bank are already expressed on the reporting scale, resulting in no need to incorporate a set of anchor items to link newly estimated item parameters to the existing scale.

Test blueprints provided the following guidelines:

- Length of the test
- Content areas to be covered and the acceptable number of items across standards within each content area or reporting category
- Acceptable range of item difficulty for the specified grade level
- Approximate number of field-test items

In grades 3–8 and 10, the NDSA ELA test includes two components, which are combined to provide an overall NDSA ELA scale score:

1. A text-based writing component in which students respond to one writing task scored in three dimensions
2. A reading, language, and listening component in which students respond to texts and multimedia content

Writing and Reading component item responses were combined to form an overall ELA score. In this document, the term *Reading* is used when referring only to the Reading tests or items; *Writing* is used when referring only to the text-based Writing task.

The NDSA uses an embedded field test (EFT) design with items placed into middling position ranges within each ELA and mathematics test. The EFT slots for spring 2021 include new field test items to replenish the broader ICCR item pool under the EFT design. EFT items are intentionally put into the middle of tests or earlier so that examinees provide the same efforts on those items as the operational items.

Table 21 displays the blueprint requirements for total test length by grade and subject. Table 22 displays the observed number of items and clusters administered during spring 2021 for each test. Blueprint requirements were satisfied at the test level for each subject and grade, except for ELA Reading grades, each of which exceeded the number of operational items by one or decreased by up to six. As many items on the ELA Reading component are associated with passages, flexibility during testing is necessary.

Table 20: Spring 2021 NDSA Item Pool by Grade and Subject

Subject	Grade	Number of Operational Items	Number of EFT Items/Clusters	Total Items/Clusters
Reading	3	417	82	499
	4	461	64	525
	5	423	75	498
	6	537	45	582
	7	529	58	587
	8	486	80	566
	10	297	7	304
Writing	3	2	—	2
	4	2	—	2
	5	2	—	2
	6	2	—	2
	7	2	—	2
	8	2	—	2
	10	2	—	2
Mathematics	3	609	48	657
	4	642	47	689
	5	531	45	576
	6	651	51	702
	7	451	40	491
	8	520	50	570
	10	675	25	700

Table 21: Blueprint Test Length by Grade and Subject

Subject	Grade	Number of Operational Items	Number of EFT Items or Clusters*	Total Test Length*
Reading	3	37–41	6–9 items OR 1 cluster	43 to 50 items OR 37 to 41 items and 1 cluster
	4	37–41	6–9 items OR	43 to 50 items OR

Subject	Grade	Number of Operational Items	Number of EFT Items or Clusters*	Total Test Length*
			1 cluster	37 to 41 items and 1 cluster
	5	37–41	6–9 items OR 1 cluster	43 to 50 items OR 37 to 41 items and 1 cluster
	6	37–41	6–9 items OR 1 cluster	43 to 50 items OR 37 to 41 items and 1 cluster
	7	37–41	6–9 items OR 1 cluster	43 to 50 items OR 37 to 41 items and 1 cluster
	8	37–41	6–9 items OR 1 cluster	43 to 50 items OR 37 to 41 items and 1 cluster
	10	37–41	6–9 items OR 1 cluster	43 to 50 items OR 37 to 41 items and 1 cluster
Writing	3	1	—	1 item
	4	1	—	1 item
	5	1	—	1 item
	6	1	—	1 item
	7	1	—	1 item
	8	1	—	1 item
	10	1	—	1 item
Mathematics	3	36–38	8 items OR 1 cluster	44–46 items OR 36–38 items and 1 cluster
	4	36–38	8 items OR 1 cluster	44–46 items OR 36–38 items and 1 cluster
	5	36–38	8 items OR 1 cluster	44–46 items OR 36–38 items and 1 cluster
	6	36–38	1–3 items AND 1 cluster	44–46 items OR 36–38 items and 1 cluster
	7	36–38	8 items OR 1 cluster	44–46 items OR 36–38 items and 1 cluster
	8	36–38	8 items OR 1 cluster	44–46 items OR 36–38 items and 1 cluster r
	10	36	8 items OR 1 cluster	44 items OR 36 items and 1 cluster

**Note: Not shown in the published blueprints (Appendix A and Appendix B)*

Table 22: Observed Spring 2021 Test Length by Grade and Subject

Subject	Grade	Number of Operational Items	Number of EFT Items or Clusters	Total Test Length
Reading	3	39–42	6–9 items OR 1 cluster	45–51 items OR 39–42 items and 1 cluster
	4	38–42	6–8 items OR 1 cluster	44–50 items OR 38–42 items and 1 cluster
	5	39–42	6–9 items OR 1 cluster	45–51 items OR 39–42 items and 1 cluster
	6	39–43	6–8 items OR 1 cluster	45–51 items OR 39–43 items and 1 cluster
	7	39–42	6–8 items OR 1 cluster	45–50 items OR 39–42 items and 1 cluster
	8	39–42	6–8 items OR 1 cluster	45–50 items OR 39–42 items and 1 cluster
	10	39–42	6 items OR 1 cluster	45–48 items OR 39–42 items and 1 cluster
Writing	3	1	—	1 item
	4	1	—	1 item
	5	1	—	1 item
	6	1	—	1 item
	7	1	—	1 item
	8	1	—	1 item
	10	1	—	1 item
Mathematics	3	36	8 items OR 1 cluster	44 items OR 36 items and 1 cluster
	4	36	8 items OR 1 cluster	44 items OR 36 items and 1 cluster
	5	36	8 items OR 1 cluster	44 items OR 36 items and 1 cluster
	6	36	1–3 items AND 1 cluster	37-39 items AND 1 cluster
	7	36	8 items OR 1 cluster	44 items OR 36 items and 1 cluster
	8	36	8 items OR 1 cluster	44 items OR 36 items and 1 cluster
	10	36	8 items OR 1 cluster	44 items OR 36 items and 1 cluster

The blueprint is designed to support reporting at multiple subdomains of the test in addition to the overall test score. Individual scores on subdomains provide information to help identify areas in which a student may have had difficulty.

Table 23 and Table 24 provide the number of operational items required in the blueprints by content strand, or subdomain, for each grade level or course. The ranges following represent an acceptable range of item counts. As many of these items on the ELA Reading component were

associated with passages, flexibility in test construction was necessary for practical reasons. These tables also provide the range of test items assessing each reporting category that appeared on the spring 2021 tests.

Table 23: Blueprint and Observed Spring 2021 Number of Test Items Assessing Each Reporting Category in ELA

Grade	Reading Informational Text	Reading Literary Text	Writing and Language	Listening*
Blueprint 3	12–14	15–17	8–10	0–3
Observed 3	12–14	16–17	10–11	1–2
Blueprint 4	12–14	15–17	8–10	0–3
Observed 4	12–14	15–17	9–11	1–2
Blueprint 5	12–14	15–17	8–10	0–3
Observed 5	12–14	15–17	10–11	1–2
Blueprint 6	14–16	14–16	8–10	0–3
Observed 6	14–16	14–15	9–11	1–3
Blueprint 7	14–16	14–16	8–10	0–3
Observed 7	14–16	14–15	10–11	1–2
Blueprint 8	14–16	14–16	8–10	0–3
Observed 8	14–16	14–15	10–11	1–1
Blueprint 10	14–16	14–16	8–10	0–3
Observed 10	14–16	14–15	10–11	1–2

**Not reported*

Table 24: Blueprint and Observed Spring 2021 Number of Test Items Assessing Each Reporting Category in Mathematics

Grade	Reporting Category			
	Operations and Algebraic Thinking	Numbers and Operations—Base Ten	Numbers and Operations—Fractions	Measurement and Data and Geometry
Blueprint 3	11–13	8–9	8–9	9–10
Observed 3	11–11	8–8	8–8	9–9
Blueprint 4	9–11	8–9	8–9	9–10
Observed 4	9–10	8–9	8–9	9–10
Blueprint 5	9–11	8–9	8–9	9–10
Observed 5	9–9	8–9	9–9	9–10

Grade	Reporting Category			
	Ratios and Proportional Relationships and Number System	Expressions and Equations	Geometry	Statistics and Probability
Blueprint 6	10–13	9–11	8–8	8–8
Observed 6	10–11	9–10	8–8	8–8
Blueprint 7	9–10	9–10	9–10	9–10
Observed 7	9–9	9–9	9–9	9–9
	Expressions and Equations and Number System	Functions	Geometry	Statistics and Probability
Blueprint 8	10–13	9–10	8–9	8–8
Observed 8	10–11	9–10	8–9	8–8
	Algebra	Functions	Geometry	Statistics and Probability
Blueprint 10	9–9	10–10	9–9	8–8
Observed 10	9–9	10–10	9–9	8–8

The summary tables show that the spring 2021 tests matched the blueprint in all cases except the WL strand for ELA, where they were not exceeded by more than one item.

In addition to information about reporting categories, the ELA Reading component and mathematics blueprints also contained target information about DOK. DOK levels are used to measure the cognitive demand of instructional objectives and assessment items. The use of DOK levels to construct the NDSA provided a greater depth and breadth of learning and also fulfilled the requirements of academic rigor required by the Every Student Succeeds Act. The DOK level described the cognitive complexity involved when engaging with an item; a higher DOK level requires greater conceptual understanding and cognitive processing by the students. It is important to note that the DOK levels are cumulative but not additive. For example, a DOK level 3 item could potentially contain DOK level 1 and 2 elements; however, DOK level 3 activity cannot be created with DOK level 1 and 2 elements.

Table 25 shows the range of the number of items by DOK level by grade and subject in the blueprint. Table 26 and Table 27 show the number of items in each DOK that appeared on the spring 2021 tests administered to students. The tables show that, in most cases, the number of items from each DOK level met the blueprint. Where the blueprint was not met, there was a maximum of a six-item difference between the blueprint and the forms. These differences occurred due to passage limits, which keep testing times down.

The Reading component blueprint also included specifications for the genres of text presented in the passages. Two main types of text were used: literary and informational.

Table 25: Blueprint Number of Items by Depth of Knowledge

Grade and Subject	DOK 1	DOK 2	DOK 3 and 4
ELA (Grades 3–8, 10)	6–9	14–24	7–13
Mathematics (Grades 3–8)	6–10	19–24	5–10
Mathematics (Grade 10)	7–9	18–23	7–9

**DOK 4 items, or writing prompts, are a larger percentage of the test points*

Table 26: Observed Spring 2021 Number of Items by Depth of Knowledge in ELA

Grade	DOK 1	DOK 2	DOK 3 and 4
3	9–15	16–22	7–13
4	8–13	16–23	8–13
5	8–13	16–24	7–13
6	7–14	16–24	7–13
7	9–12	17–24	8–13
8	7–13	16–23	7–13
10	7–10	18–25	7–13

**DOK 4 items, or writing prompts, are a larger percentage of the test points.*

Table 27: Observed Spring 2021 Number of Items by Depth of Knowledge in Mathematics

Grade	DOK 1	DOK 2	DOK 3
3	7–9	21–23	5–7
4	7–9	21–23	5–7
5	7–9	20–23	5–8
6	7–9	21–22	5–7
7	7–10	20–23	5–7
8	7–10	20–22	6–7
10	7–9	19–22	7–9

4.2 Test Construction

As discussed in Volume 1, the NDDPI made the decision to change the format of the NDSA from a fixed form to a computerized adaptive test for the current test administration. In fall 2020, CAI psychometricians and content experts worked with NDDPI content specialists and leadership to build item pools for the spring 2021 administration. NDSA test construction used a structured test

construction plan, explicit blueprints, and active collaborative participation from all parties. ELA and mathematics assessments employ computer adaptive testing that draws from item pools. For more information about CAI's adaptive algorithm, see Appendix M.

The 2021 NDSA test item pools were built by CAI test developers to support exact match to the detailed test blueprints and target distributions of item difficulty and test information. Operational items were selected to fulfill the blueprint for that grade. The subsequent sections outline the roles and responsibilities of the participants, test construction process, materials used, and sample statistical and graphical summaries used during the review process.

As mentioned, the blueprint describes the content to be covered, the DOK with which it will be covered, the type of items that will measure the constructs, and other content-relevant aspects of the test. The psychometric considerations, ensuring that students will receive scores of similar precision, include the following:

- A reasonable range of item difficulties was included.
- *p*-values for items were reasonable and within specified bounds (> 5% and < 95%).
- Biserial correlations were reasonable and within specified bounds (> 0.20 for correct response and < 0 for incorrect responses).
- For all items, item response theory (IRT) *a*-parameters were reasonable and greater than 0.4.
- For all items, IRT *b*-parameters were reasonable.
- For multiple choice items, IRT *c*-parameters were less than 0.40.

More information about *p*-values, biserial correlations, and IRT parameters can be found in Volume 1. The details on calibration, equating, and scoring of the NDSA can also be found in Volume 1.

4.3 Roles and Responsibilities

4.3.1 Role of the CAI Content Team

CAI ELA and mathematics content teams were responsible for the initial item pool construction and subsequent revisions. CAI content teams performed the following tasks:

- Selection of the operational items
- Revision of the operational item sets according to feedback from senior CAI content staff
- Revision of the operational item sets included according to feedback from the CAI technical team
- Revision of the operational item sets according to feedback from NDDPI
- Assistance in the generation of materials for NDDPI review

- Revision of the item pools to incorporate feedback from NDDPI

4.3.2 Role of the CAI Technical Team

The CAI technical team, which includes psychometricians and statistical support associates, prepares the item bank by updating ITS with current item statistics and provides test construction training to the internal content team. During test construction, at least one psychometrician facilitates each content area. The technical team performs the following tasks:

- Preparing item bank statistics and updating CAI's ITS
- Creating the master data sheets (MDS) for each grade and subject
- Providing feedback on the statistical properties of initial item selections
- Providing feedback on the statistical properties of each subsequent item selection
- Creating statistical summary and materials for NDDPI review

4.3.3 State Content Specialists and Reviewers

NDDPI invited teachers from the field to review the proposed item pools. The review process involved use of content and blueprint guidelines in addition to the statistical guidelines. NDDPI leadership was also involved in the review process and made the final decision for approval. When evaluating any given item pools, leadership considered the diversity of topics, projected level of difficulty, statistical summaries, adherence to blueprint, overall challenge to the examinees, and acceptability of test content to the North Dakota public. NDDPI was given the opportunity to approve proposed item pools or return them with comments to CAI's content and psychometric teams for further revision.

Appendix L summarizes the characteristics of the committees who reviewed, and ultimately approved, the item pools.

5. REFERENCES

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014*). Standards for educational and psychological testing*.

Calisir, F., & Gurel, Z. (2003). Influence of text structure and prior knowledge of the learner on reading comprehension, browsing and perceived control. *Computers in Human Behavior, 19*(2), 135–145.

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.

Egan, K. L., Schneider, M. C., & Ferrara, S. (2012). Performance level descriptors: History, practice, and a proposed framework. In G. J. Cizek (Ed.) *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 79-106). New York: Routledge.

Fisher, D., Frey, N., & Lapp, D. (2012). *Text complexity: Raising rigor in reading*. Newark, DE.: International Reading Association.

Freebody, P., & Anderson, R. C. (1983). Effects on Text Comprehension of Differing Proportions and Locations of Difficult Vocabulary. *Journal of Reading Behavior*, *15*(3), 19–39.

Gillioz, C., Gygax, P., & Tapiero, I. (2012). Individual differences and emotional inferences during reading comprehension. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 66*(4), 239–250.

Kucer, S. B. (2010). Going beyond the author: What retellings tell us about comprehending narrative and expository texts. *Literacy, 45*(2), 62–69.

Long, D. L., & De Ley, L. (2000). Implicit causality and discourse focus: The interaction of text and reader characteristics in pronoun resolution. *Journal of Memory and Language, 42*(4), 545–570.

McConaughy, S. (1985). Good and poor readers' comprehension of story structure across different input and output modalities. *Reading Research Quarterly, 20*(2), 219–232. doi:10.2307/747757

Petit, M. & Hess, K. (2008). *Applying Webb's Depth of Knowledge and NAEP Levels of Complexity in Mathematics.* Retrieved from the National Center for Assessment website: https://www.nciea.org/sites/default/files/publications/DOKmath_KH08.pdf

Rapp, D. N., & Mensink, M. C. (2011). Focusing effects from online and offline reading tasks. In M. T. McCrudden, J. P. Magliano, & G. Schraw (Eds.), *Text relevance and learning from text* (pp. 141–164). Charlotte, NC, US: IAP Information Age Publishing.

Rich, S. S., & Taylor, H. A. (2000). Not all narrative shifts function equally. *Memory & Cognition, 28*(7), 1257–1266.

Riding, R. J., & Taylor, E. M. (1976). Imagery performance and prose comprehension in seven-year-old children. *Educational Studies, 2*(1), 21–2.

Rommers, J., Dijkstra, T., & Bastiaansen, M. (2013). Context-dependent semantic processing in the human brain: Evidence from idiom comprehension. *Journal of Cognitive Neuroscience, 25*(5), 762–776.

Sadoski, M., Goetz, E. T., & Fritz, J. B. (1993). A causal model of sentence recall: Effects of familiarity, concreteness, comprehensibility, and interestingness. *Journal of Reading Behavior, 25*(1), 5–16.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The Percentage of Words Known in a Text and Reading Comprehension. *Modern Language Journal*, *95*(1), 26–43.

Sparks, J. R., & Rapp, D. N. (2011). Readers reliance on source credibility in the service of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(1), 230–247.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved February 15, 2012, from http://www.cehd.umn.edu/NCEO/onlinepubs/Synthesis44.html.

Webb, N. L. (2002). *Alignment study in language arts, mathematics, science, and social studies of state standards and assessments for four states*. Washington, DC: Council of Chief State School Officers.