New Hampshire Statewide Assessment System

2018–2019

Volume 2, Part 1 (ELA and Mathematics) Test Development

New Hampshire Department of Education
Serving New Hampshire's Education Community

TABLE OF CONTENTS

LIST OF EXHIBITS

LIST OF TABLES

LIST OF APPENDICES

1. INTRODUCTION

The AIRCore English language arts (ELA) and mathematics item bank is written to measure career- and college-readiness standards as embodied in the Common Core State Standards (CCSS). 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 because 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. AIRCore item specifications were developed in partnership with the state of Utah, and the item specifications began as a joint endeavor. At the outset, the AIRCore item specifications matched the Utah Student Assessment of Growth and Excellence (SAGE) item specifications. Utah SAGE received the approval of peer reviewers, validating the quality and alignment of the specifications to the career- and college-ready standards. Over time, the specifications have been updated to incorporate an expanding set of potential interactions and item types. An expanding pool of states adopted AIRCore as a component of their item pool or, in some cases, the entire basis of their tests. As described in the following sections, each AIRCore item has been through a series of stakeholder reviews in one or more participating states.

1.1 CLAIM STRUCTURE

The assessment is designed to measure career and college 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 texts.
- Students can read closely and analytically to comprehend a range of increasingly complex 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 AIRCore 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.

Exhibit A: AIRCore Mathematics Procedural Categories Forming the Basis of Subclaims by 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

1.2 UNDERLYING PRINCIPLES GUIDING DEVELOPMENT

The AIRCore item bank was established using a highly structured, evidence-centered design. The process began with detailed item specifications. The 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 AIR's test delivery platform, which has received Web Content Accessibility Guidelines (WCAG) 2.0 AA certification, offers a wide array of accessibility tools, and is compatible with most assistive technologies.

Item development supported the goal of 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 the item.

AIR 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 ELA and mathematics item development process that supports the validity of the claims that AIRCore tests are designed to support
2. An overview of the ELA and mathematics 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 New Hampshire Statewide Assessment System (NH SAS) for ELA and mathematics, including the blueprint design and test construction

2. ITEM DEVELOPMENT PROCESS THAT SUPPORTS VALIDITY OF CLAIMS

2.1 OVERVIEW

AIR developed the AIRCore ELA and mathematics item banks using a rigorous, structured process that engaged stakeholders at critical junctures. This process was managed by AIR'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 AIR 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 below.

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	Training in language accessibility, bias, and sensitivity helps item	—

	Supports Alignment to the Standards	Reduces Construct-Irrelevant Variance Through Universal Design	Expands Access Through Linguistic and Other Supports
	and specifications. Teaches item writers about selection of item types for measurement and accessibility.	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 developed in collaboration between content experts in the Utah State Board of Education and AIR content experts. The specifications were used to develop both the SAGE pool and the AIRCore pool. 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 CCSS ELA standards and accompanying material. Exhibit C provides sample passage specifications.

Exhibit C: Sample Passage Specifications

Difficulty Factor	Passage Metric Description	Grade-Level Details (sample for grades 9–10)	Research-Based Evidence
Levels of Meaning in Literature	1. Single, concrete interpretation with few generalizations necessary. 2. Some themes are not explicitly stated. 3. There are multiple, successively abstract or general levels of meaning; key theme or themes are implied.	1. a. Characters are static, and characteristics are explicitly stated. b. Setting is used as an aesthetic enhancement, not as a way to convey meaning. c. Mood and tone are used to enhance the setting of the story but are not critical in conveying the meaning or theme. d. Actions have straightforward meanings and clear, immediate effects. e. Symbols are straightforward, common, and closely linked to their meanings, both in terms of proximity and explanatory language. 2. a. Characters are dynamic, and a single character may have multiple motives. b. Characteristics are implied through clear action or dialogue. c. Setting serves to underscore the theme and conveys mood or tone, which supports understanding of the explicit theme. d. Actions have straightforward, explicit meanings, but the effects are not fully realized until later in the passage. e. Symbols are straightforward and common but may not be supported by explanation or elaboration (e.g., children's bare feet symbolize poverty, which is not explained but can be deduced through context). f. There may be some simple analogies or allusions to other works. 3. a. Characters are complex with multiple motives and/or inner conflicts.	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 (e.g., descriptions of specific traits as being either positive or negative) (Rapp & Mensink, 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

Difficulty Factor	Passage Metric Description	Grade-Level Details (sample for grades 9–10)	Research-Based Evidence
		b. Characterization is implied through subtle actions, others' reactions, and oblique dialogue. c. The setting is used to reveal the theme. d. Setting conveys mood or tone, which is crucial to understanding the implicit theme. e. Reader may need to understand historical context to fully comprehend text. f. Actions have subtle and/or complex meanings, the effects of which may not be immediately realized. g. Symbols are complex, uncommon, and/or make assumptions about students' historical, scientific, or literary knowledge. h. There may be complex analogies or allusions to other works.	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, 1985). The use of figurative language and meanings also increases the difficulty of a text. (Rommers, Dijkstra, & Bastiaansen, 2013). It is easier to understand texts when their words stand for their literal meanings. Figurative language such as satire, irony, and allusions is more difficult to interpret than figurative language like imagery or metaphors (Fisher, Frey, & Lapp, 2012).
Structure	1. There is a 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. Story is presented in a straightforward fashion without any shifts in time or narrator. At this grade level, this includes significant digression into details and setting, as long as the chronology is consistent. 2. a. Narrator shifts with a clear signal that it is doing so. b. Includes simple chronology shifts, such as clearly introduced flashbacks or memories. c. Structure varies with a mixture of prose and verse, OR progresses in a nonlinear fashion. 3. a. The narrator shifts but may not give a clear signal that it is doing so. b. Includes complex chronology shifts, such as flashbacks or memories.	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 retell stories, with more flashbacks and more narrator changes throughout a story

Difficulty Factor	Passage Metric Description	Grade-Level Details (sample for grades 9–10)	Research-Based Evidence
		c. Structure varies with a mixture of prose and verse, OR progresses in a nonlinear fashion.	compounding this effect (Kucer, 2010).
Language	1. Simple, common word choice, explicit and literal use. 2. May include unfamiliar vocabulary; abstract meaning; or figurative, ironic, or sarcastic use. 3. Generally dense, using figurative or purposefully ambiguous, often unfamiliar language	1. Uses high-frequency, grade-appropriate vocabulary that relies on denotative meaning. Minimal use of literary devices. Syntax is clear and consistent. 2. a. Uses unfamiliar, above–grade-level words. b. Uses at-grade-level words with intended multiple connotations in order to convey multiple meanings. c. Uses common colloquialisms and/or simple dialect. d. Uses simple literary devices and figurative language. 3. a. Words are unfamiliar, archaic, or academic. b. Some words cannot be fully comprehended with context clues. c. Uses authentic, complex dialect, colloquialisms, and/or vernacular which may make assumptions about students' prior experience. d. Uses complex or abstract figurative language or literary devices.	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).
Total Score			
Key	1. Scores below 5 indicate easy content. 2. Scores from 5–8 indicate medium-difficulty content. 3. Scores from 9–12 indicate difficult content.		

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

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

- Practice Cluster 1: Use Mathematics to Solve Problems
  - MP1—Make sense of problems and persevere in solving them.
  - MP4—Model with mathematics.
  - MP5—Use appropriate tools strategically.
- Practice Cluster 2: Use Mathematical Reasoning
  - MP2—Reason abstractly and quantitatively.
  - MP3—Construct viable arguments and critique the reasoning of others.
  - MP6—Attend to precision.
- Practice Cluster 3: Use Characteristics of Problems to Generalize
  - 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, or acceptable shapes for geometry standards.
- **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 being 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 items, 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 PC or PCs a particular standard could be aligned: PC1, PC2, PC3, or none.
- **Depth of Knowledge.** The task demands of each standard can be classified as DOK 1, DOK 2, or DOK 3.
- **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.
- **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, and reflection of the performance 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 limited to 2, 3, 4, 5, 6, 8, 10, 12, 100. *Benchmarks 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.	• Multi-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 • Multi-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 performance level. Conversely, if a student is fluently comparing	x	x	x	

			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.				

	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 [include at least two more pairs of fractions]	Hard	1, 2	Matching response	This is a DOK 2 because students are comparing fractions using <, >, or =. 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, it is noted whether evidence-based selected-response (two-part), extended-response, hot text, multiple-choice, multiple-select,

and/or short answer (to be scored automatically with our *proposition scorer*) items may be used, and if so, how.

- **DOK Demands.** This section is divided 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," the acceptable response mechanisms to hot text were limited, wherein the student selects the evidence in the text itself, and multi-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.

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.	*Roll of Thunder, Hear My Cry*

			Easy Difficulty: The support for the inference stated in the question is explicitly provided in the text.	
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 the 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 AIRCore items have at least a bachelor's degree, and many bring teaching experience. All item writers are trained in the following disciplines:

- The principles of universal design
- The appropriate use of item types
- The AIRCore specifications

Key materials are shown in Appendix A. These include

- AIR's Language Accessibility, Bias, and Sensitivity 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 A, B, and C, respectively.

2.4 INTERNAL REVIEW

AIRCore's test development structure uses 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 before client review and provide training and feedback for all content-area team members.

AIRCore items go through 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 item review checklist that our test developers use is included in Appendix B. The AIRCore internal review cycle includes the following phases:

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

2.4.1 Preliminary Review

Team leads or senior content staff conduct Preliminary Review. 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 convey 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 is

- as succinct 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 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 from both the perspective of potential clients as well as 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 items is correct. For mathematics items, editors perform all calculations to ensure accuracy.

Third, editors review all material for fairness and language accessibility issues.

Finally, editors confirm that items reflect the accepted guidelines for good item construction. In all items, they look for language that is simple, direct, free of ambiguity, and 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. They also ensure 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 AIRCore item arrives at Senior Review, both content reviewers and editors have thoroughly vetted it. 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 AIRCore 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 AIRCore item bank, state content experts review any eligible items before committee review. At this stage in the review process, clients can request edits, such as wording edits, scoring edits, alignment changes, or DOK updates. An AIR

director for mathematics or ELA reviews all client-requested edits in light of the AIRCore 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 (because such edits risk altering the function of calibrated items), and clients can simply select the items from the available item bank to present to the committee.

2.5.2 Content Advisory Committee Reviews

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 that follows).

A summary of the committee meetings appears in Exhibit F, with further details about the participants in Appendix C.

Exhibit F: Summary of 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

Location	Year	Number of Committee Members	Number of Items Reviewed
West Virginia	2018	24	317
Wyoming	2018	36	503

2.5.3 Language Accessibility, Bias, and Sensitivity Committee Reviews

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 in Appendix D.

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 translations or markups 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, AIRCore items are tagged with Spanish translations and text-to-speech, including Spanish text-to-speech.

2.6 FIELD TESTING

AIRCore items were field tested embedded in operational, summative accountability assessments in participating states. AIR's field-testing process is described in detail in Volume 1, Section 5.1.

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 below.

2.7.1 Key Verification

Key verification is a simple process by which a frequency table of response frequencies and the scores that they received is created. Qualified content staff review them to ensure that all correct responses, and only correct responses, receive a score.

2.7.2 Rubric Validation

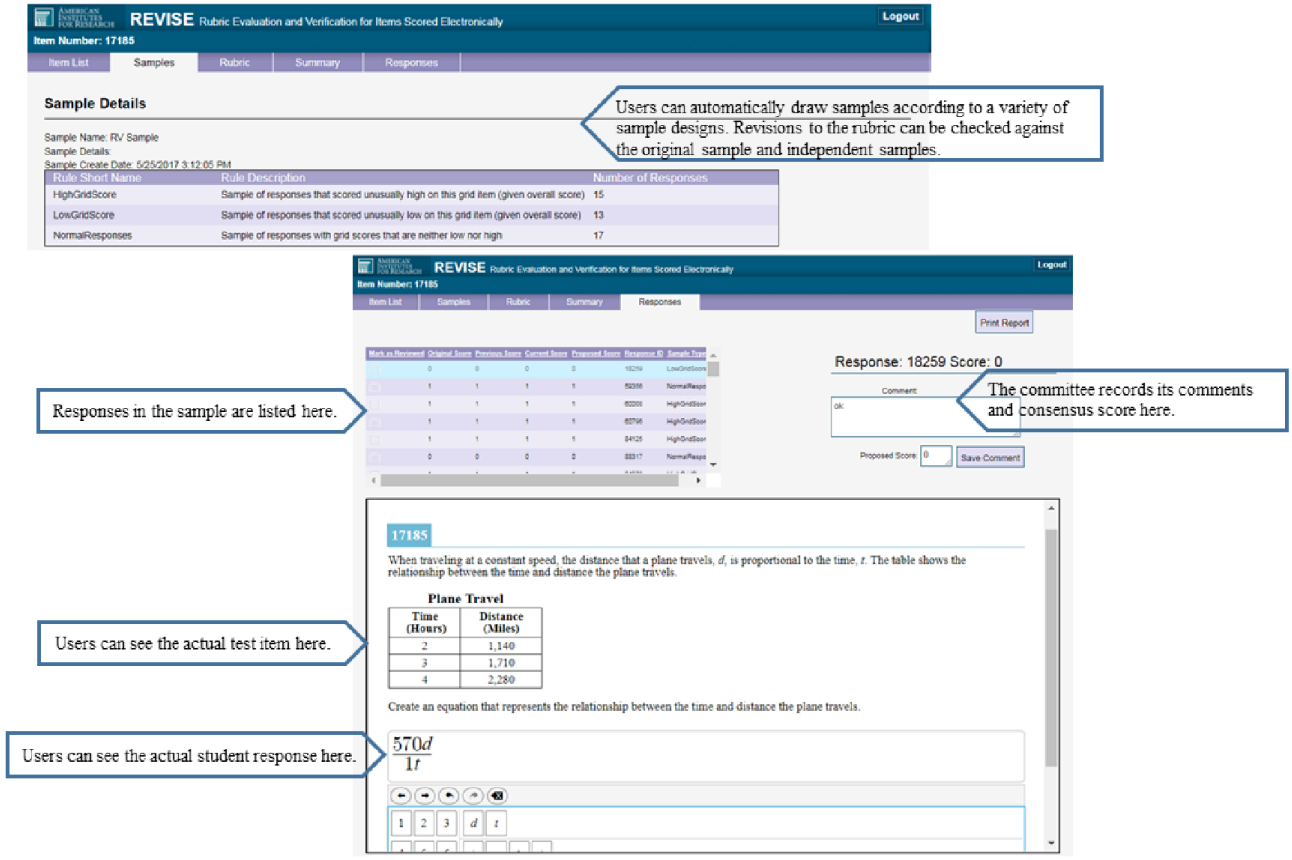
More complex selected-response items, as well as machine-scored, constructed-response items, undergo rubric validation, which occurs in two phases. During the first phase, AIR content experts draw one or more samples to identify anomalous or unforeseen responses and ensure that they are scored correctly.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

The second phase of rubric validation involves state content experts. During this phase, a fresh sample of responses are 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 called *REVISE*. Items are reviewed as the students saw them, along with the student's response. 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 of 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 handscoring 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 handscoring. Handscoring 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 E.

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

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. AIRCORE ITEM BANK SUMMARY

The AIRCore item bank is robust and has been constructed explicitly to support multiple statewide assessment programs. As described above, AIRCore items were written to the Common Core State Standards, and the bank is occasionally augmented with items measuring some state-specific standards. The AIRCore 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 AIRCore 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 AIRCore continues to grow as AIR 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 in the following sections.

3.1 CURRENT COMPOSITION OF THE ITEM BANK

Table 1 and Table 2 list the ELA and mathematics item types and briefly describe each. Examples of various item types can be found in Appendix G.

Table 1: ELA Item Types and Descriptions

Response Type	Description
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.
Editing Task Choice (ETC)	Student identifies an incorrect word or phrase and chooses the replacement from a number of options.
Grid (GI)	Student selects words, phrases, or images and uses the drag-and-drop feature to place them into a graphic organizer.
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.
Matching (MI)	Student checks a box to indicate if information from a column header matches information from a row.
Multiple Choice (MC)	Student selects one correct answer from a number of options.
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 Select (MS)	Student selects all correct answers from a number of options.
Natural Language (NL)	Student uses the keyboard to enter a response into a text field.

**Note: the abbreviations correlate to the attributes used in AIR's Item Tracking System.*

Table 2: Mathematics Item Types and Descriptions

Response Type	Description
Equation (EQ)	Student uses a toolbar with a variety of mathematical symbols to create a response.
Editing Task Choice (ETC)	Student identifies an incorrect word or phrase and chooses the replacement from a number of options.
Grid (GI)	Student selects words, phrases, or images and uses the drag-and-drop feature to place them into a graphic organizer.
Matching (MI)	Student checks a box to indicate if information from a column header matches information from a row.
Multiple Choice (MC)	Student selects one correct answer from a number of options.
Multiple Select (MS)	Student selects all correct answers from a number of options.
Table (TI)	Student types numeric values into a given table.

**Note: the abbreviations correlate to the attributes used in AIR's Item Tracking System.*

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

Table 3: AIRCore ELA Spring 2019 Operational and Field-Test Item Pool

Grade	Total Number of Items	Number of Writing Prompts
3	455	6
4	506	11
5	497	13
6	584	13
7	591	16
8	542	14
9	407	11
10	421	12
11	344	14
Total	***4,347***	***110***

Table 4: AIRCore ELA Spring 2019 Operational Item Pool

Grade	Number of Total OP Items
3	325
4	397
5	361
6	455
7	472
8	439
9	307
10	311
11	304
Total	***3,371***

Table 5: AIRCore ELA Spring 2019 Field-Test Item Pool

Grade	Number of Total FT Items
3	130
4	109
5	136
6	129
7	119
8	103
9	100
10	110
11	40
Total	***976***

Table 6: AIRCore ELA Spring 2019 Item Counts by Grade and Reporting Category

Grade	Reading Informational Text	Reading Literary Text	Writing and Language	Speaking and Listening	Grand Total
3	202	149	97	7	455
4	208	156	135	7	506
5	186	165	135	11	497
6	255	178	135	16	584
7	234	206	141	10	591
8	240	159	137	6	542
9	146	143	110	8	407
10	171	128	117	5	421
11	159	79	102	4	344
Total	***1,801***	***1,363***	***1,109***	***74***	***4,347***

Table 7: AIRCore ELA Spring 2019 Item Counts by Grade and DOK

Grade	DOK 1	DOK 2	DOK 3	DOK 4	Grand Total
3	85	304	60	6	455
4	95	332	68	11	506
5	86	327	71	13	497
6	87	383	100	14	584
7	80	379	116	16	591
8	80	343	105	14	542
9	59	265	72	11	407
10	69	256	83	13	421
11	52	196	82	14	344
Total	***693***	***2,785***	***757***	***112***	***4,347***

Table 8: AIRCore ELA Spring 2019 Item Counts by Grade and Item Type

Grade	Item Type	Number of Items
3	Editing Task Choice	49
	Extended Response	6
	Hot Text	41
	Multiple Choice	317
	Multiple Choice, Hot Text	1
	Multiple Choice, Multiple Select	3
	Table Match	13
	Multiple Select	24
	Multiple Select, Hot Text	1
	Total	***455***

Grade	Item Type	Number of Items
4	Editing Task Choice	61
	Extended Response	6
	Hot Text	43
	Multiple Choice	332
	Multiple Choice, Hot Text	5
	Multiple Choice, Multiple Select	3
	Table Match	10
	Table Match, External Copy Inline, Text Entry	1
	Multiple Select	38
	Natural Language	2
	Text Entry	5
	Total	***506***
5	Editing Task Choice	64
	Editing Task Choice, Multiple Choice	1
	Extended Response	6
	Grid	1
	Hot Text	51
	Multiple Choice	293
	Multiple Choice, Hot Text	4
	Multiple Choice, Multiple Select	6
	Table Match	20
	Multiple Select	43
	Natural Language	1
	Text Entry	7
	Total	***497***
6	Editing Task Choice	64
	Editing Task Choice, Table Match, Multiple Select, External Copy Block, Text Entry	1
	Extended Response	6
	Hot Text	39
	Multiple Choice	395
	Multiple Choice, Hot Text	9
	Multiple Choice, Multiple Select	4
	Table Match	10
	Multiple Select	48
	Natural Language	1
	Text Entry	7
	Total	***584***

Grade	Item Type	Number of Items
7	Editing Task Choice	63
	Editing Task Choice, Multiple Choice	1
	Extended Response	6
	Hot Text	42
	Multiple Choice	355
	Multiple Choice, Hot Text	1
	Multiple Choice, Multiple Select	23
	Table Match	6
	Multiple Select	81
	Natural Language	3
	Text Entry	10
	Total	***591***
8	Editing Task Choice	59
	Extended Response	6
	Hot Text	38
	Multiple Choice	375
	Multiple Choice, Hot Text	4
	Multiple Choice, Multiple Select	3
	Table Match	7
	Multiple Select	42
	Text Entry	8
	Total	***542***
9	Editing Task Choice	57
	Extended Response	3
	GI	1
	Hot Text	43
	Multiple Choice	260
	Multiple Choice, Multiple Select	6
	Table Match	2
	Multiple Select	25
	Natural Language	2
	Text Entry	8
	Total	***407***

Grade	Item Type	Number of Items
10	Editing Task Choice	64
	Editing Task Choice, Multiple Choice	1
	Editing Task Choice, Multiple Choice, Multiple Select, External Copy Block, Text Entry	1
	Extended Response	2
	Hot Text	38
	Multiple Choice	270
	Multiple Choice, Hot Text	3
	Multiple Choice, Multiple Select	2
	Table Match	1
	Multiple Select	28
	Natural Language	1
	Text Entry	10
	Total	***421***
11	Editing Task Choice	49
	Hot Text	34
	Multiple Choice	211
	Multiple Choice, Multiple Select	5
	Table Match	1
	Multiple Select	27
	Natural Language	3
	Text Entry	14
	Total	***344***
All	***Grand Total***	***4,347***

Table 9: AIRCore Mathematics Spring 2019 Operational and Field-Test Item Pool

Grade	Total Number of Items
3	616
4	654
5	524
6	670
7	467
8	532
HS	1,349
Total	***4,812***

Table 10: AIRCore Mathematics Spring 2019 Operational Item Pool

Grade	Number of Spring 2019 OP Items
3	480
4	496
5	409
6	473
7	365
8	415
HS	1,096
Total	***3,734***

Table 11: AIRCore Mathematics Spring 2019 Field-Test Item Pool

Grade	Number of Spring 2019 FT Items
3	136
4	158
5	115

Grade	Number of Spring 2019 FT Items
6	197
7	102
8	117
HS	253
Total	***1,078***

Table 12: AIRCore Mathematics Spring 2019 Item Counts by Grade and Reporting Category

Grade	Reporting Category	Number of Items
3	Geometry	46
	Measurement and Data	121
	Number and Operations—Fractions	160
	Number and Operations in Base Ten	108
	Operations and Algebraic Thinking	181
	Total	***616***
4	Geometry	65
	Measurement and Data	99
	Number and Operations—Fractions	193
	Number and Operations in Base Ten	183
	Operations and Algebraic Thinking	114
	Total	***654***
5	Geometry	58
	Measurement and Data	72
	Number and Operations—Fractions	159
	Number and Operations in Base Ten	148
	Operations and Algebraic Thinking	87
	Total	***524***
6	Expressions and Equations	201
	Geometry	75
	Ratios and Proportional Relationships	165
	Statistics and Probability	71
	The Number System	158
	Total	***670***

Grade	Reporting Category	Number of Items
7	Expressions and Equations	84
	Geometry	101
	Ratios and Proportional Relationships	93
	Statistics and Probability	101
	The Number System	88
	Total	***467***
8	Expressions and Equations	157
	Functions	108
	Geometry	132
	Statistics and Probability	73
	The Number System	62
	Total	***532***
HS	Algebra	317
	Functions	359
	Geometry	416
	Number and Quantity	82
	Statistics and Probability	175
	Total	***1,349***
All	***Grand Total***	***4,812***

Table 13: AIRCore Mathematics Spring 2019 Item Counts by Grade and DOK

Grade	DOK 1	DOK 2	DOK 3	Total
3	145	389	82	616
4	154	421	79	654
5	105	351	68	524
6	168	433	69	670
7	79	304	84	467
8	121	310	101	532
HS	179	995	175	1,349
Total	***951***	***3,203***	***658***	***4,812***

Table 14: AIRCore Mathematics Spring 2019 Item Counts by Item Type

Grade	Item Type	Number of Items
3	Equation	332
	Editing Task Choice	1
	Grid	84
	Multiple Choice	125
	Table Match	11
	Multiple Select	48
	Table Input	15
	Total	***616***
4	Equation	349
	Editing Task Choice	5
	Editing Task Choice, Equation	1
	Grid	56
	Multiple Choice	101
	Multiple Choice, Equation	1
	Table Match	31
	Multiple Select	95
	Table Input	15
	Total	***654***
5	Equation	322
	Editing Task Choice	5
	Grid	27
	Multiple Choice	103
	Table Match	12
	Multiple Select	44
	Table Input	11
	Total	***524***
6	Equation	342
	Editing Task Choice	1
	Grid	42
	Multiple Choice	184
	Table Match	13
	Multiple Select	56
	Table Input	32
	Total	***670***
7	Equation	285
	Editing Task Choice, Equation	1

Grade	Item Type	Number of Items
	Grid	38
	Multiple Choice	112
	Table Match	10
	Multiple Select	18
	Table Input	3
	Total	***467***
8	Equation	234
	Editing Task Choice	4
	Grid	52
	Grid, Equation	1
	Multiple Choice	178
	Multiple Choice, Equation	1
	Table Match	9
	Multiple Select	45
	Table Input	8
	Total	***532***
HS	Equation	563
	Editing Task Choice	18
	Editing Task Choice, Equation	5
	Editing Task Choice, Multiple Choice	1
	Grid	64
	Grid, Equation	1
	Hot Text	36
	Multiple Choice	565
	Multiple Choice, Equation	2
	Table Match	15
	Multiple Select	70
	Multiple Select, Equation	1
	Table Input	8
	Total	***1,349***
All	***Grand Total***	***4,812***

Table 15: AIRCore Cluster Item Counts

Grade	Third-Generation Performance Tasks
ELA	
3	—
4	1
5	—
6	1
7	—
8	—
HS	1
Mathematics	
3	3
4	3
5	3
6	4
7	5
8	4
HS	2

3.2 STRATEGY FOR POOL EVALUATION AND REPLENISHMENT

AIR 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. AIR 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 (EFT) slots. AIR's field-testing design is described in detail in Volume 1, Section 5.1. 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. Characteristics of released items to be replaced
2. Characteristics of items overused in adaptive programs
3. Tabulations of content coverage and ranges of difficulty to identify gaps in the pool

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. NH SAS TEST CONSTRUCTION

Using AIRCore as the source of items for the NH SAS in ELA and mathematics, tests in New Hampshire were constructed to meet the state-specific test blueprints that were written to align with the New Hampshire College and Career Ready Standards (NH CCRS). Because the AIRCore item bank is large and contains an array of item types, the tests could be uniquely developed by drawing from the pool of available AIRCore items. The construction of test item pools for the online ELA and mathematics NH SAS is a process that requires both expert judgement from content experts and 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 NH SAS is designed to support the claims described in the outset of this volume. AIR worked closely with the New Hampshire Department of Education (NHDOE) to create blueprints that guided the development process for the NH SAS. Blueprints were designed to meet the following objectives:

- Full coverage of the breadth and depth of the NH CCRS
- 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 TEST BLUEPRINTS

Test blueprints provide 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
- Approximate number of field-test items, if applicable

The NH SAS ELA assessment includes two components, which are combined to provide overall NH SAS ELA scale scores:

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 technical report, the term *Reading* is used when referring only to the Reading test component or items; *Writing* is used when referring only to the text-based Writing task.

4.1.1 ELA Blueprints

The detailed blueprints developed for English language arts grades 3–8 are provided in Appendix H. The blueprints are organized by strand and specify the number of items required for each reporting category, ensuring that the test contains enough items at that category to elicit enough information from the student to justify strand-level scores.

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
- Three to five 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, included in Table 27.

Because the AIRCore item bank offers a range of item types to assess all the standards described above, each item pool constructed fulfills the NH SAS 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.

While tests are not timed, NHDOE published estimated testing times for the NH SAS ELA Reading component within the blueprint, represented in Table 16. To estimate these Reading times, AIR analyzed the average testing time for students where AIRCore items were field tested. The average page time per item for reading literature and informational passages were computed, then multiplied 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. As an additional aid to teachers and administrators, the NHDOE provided a broad overview of estimated testing times that included both ELA components (available at https://nh.portal.airast.org/resources/general-information-resources/), also shown in Table 16. Observed testing times in Table 17 represent the first year of test administration for the NH SAS. All ELA Reading times are around or less than the estimates. Observed NH SAS testing times will be continually monitored for abnormalities over future test administrations.

Table 16: Estimated ELA Testing Times by Grade

Subject	Grade	85th Percentile Testing Time from Blueprints (minutes)	Estimated Testing Time Overview (minutes)
Reading	3	125	120
	4	99	120
	5	107	120
	6	102	120
	7	107	120
	8	85	120
Writing	3	—	120
	4	—	120
	5	—	120
	6	—	120
	7	—	120
	8	—	120

Table 17: Observed Spring 2019 ELA Testing Times by Grade

Subject	Grade	85th Percentile Testing Time (minutes)
Reading	3	111
	4	100
	5	103
	6	95
	7	92
	8	79
Writing	3	115
	4	123
	5	114
	6	104
	7	97
	8	89

4.1.2 Mathematics Blueprints

The blueprints developed for mathematics grades 3–8 are shown in Appendix I. 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 while maintaining a structure that emphasizes some reporting categories over others.

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 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 mathematics blueprints also contain item ranges for DOK, shown in Table 27. These item ranges ensure that all students are exposed to varying levels of cognitive complexity while still providing some flexibility during test construction.

The AIRCore 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, AIR and NHDOE 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.

The published estimated testing times for mathematics are shown in Table 18. To estimate these times, AIR first looked at the average testing time of students on typical AIRCore mathematics items. In general, across all grades, students spent more time on machine-scored constructed-response items than on selected-response items. Using the proportion of each specific item type with regard to the item type category within the AIRCore item bank, the average time spent on Selected-Response and Machine-Scored Constructed-Response items was calculated, given the composition of the item bank. Based on these averages and the range of number of items per test, the rough estimates mathematics testing times provided in Table 18 were determined. The observed testing times in Table 19 represent the first year of test administration for the NH SAS and are somewhat less than the projected times. Observed NH SAS testing times will be continually monitored for abnormalities over future test administrations.

Table 18: Estimated Mathematics Testing Times by Grade

Grade	Estimated Testing Time Overview (min)
G3	120
G4	120

Grade	Estimated Testing Time Overview (min)
G5	120
G6	120
G7	120
G8	120

Table 19: Observed Spring 2019 Mathematics Testing Times by Grade

Grade	85th Percentile Testing Time (min)
G3	103
G4	96
G5	100
G6	102
G7	90
G8	84

4.1.3 Overview of NH SAS 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.

The NH SAS 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 2019 include new field-test items to replenish the broader AIRCore 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 20 shows the number of operational and EFT items available in the NH SAS item pool for administration during spring 2019 testing. Table 21 displays the blueprint requirements for operational items by grade and subject. Table 22 displays the observed number of items administered during spring 2019 for each subject and grade. Blueprint requirements are satisfied at the total test level.

Table 20: Spring 2019 NH SAS Item Pool by Grade and Subject

Subject	Grade	Number of Operational Items	Number of EFT Items	Total Items
Reading	3	300	101	401
	4	353	78	431
	5	306	78	384
	6	417	95	512
	7	429	76	505
	8	385	79	464
Mathematics	3	479	96	575
	4	495	109	604
	5	409	83	492
	6	468	151	619
	7	338	63	401
	8	411	91	502

Table 21: Blueprint Test Length by Grade and Subject

Subject	Grades	Number of Operational Items	Number of EFT Items*	Total Test Length*
Reading	3–8	35–37	7–9 items OR 0–2 items and 1 cluster	41–46
Writing	3–8	1	—	1
Mathematics	3–5, 7–8	34	8 items OR 1 cluster	42 items OR 34 items and 1 cluster
	6	34	8 items OR 3 items and 1 cluster	42 items OR 37 items and 1 cluster

**Not included in the published blueprints (Appendix H and Appendix I)*

Table 22: Observed Spring 2019 Test Length by Grade and Subject

Subject	Grade	Number of Operational Items	Number of Linking/EFT Items	Total Test Length
Reading	3	35–39	7–9	42–48
	4	35–39	8	35–47
	5	35–37	8	35–45

Subject	Grade	Number of Operational Items	Number of Linking/EFT Items	Total Test Length
	6	35–38	1 cluster OR 8 Items	35–46
	7	35–38	7–8	42–46
	8	35–37	7–9	42–46
Writing	3–8	1	—	1
Mathematics	3	34	8 items OR 1 cluster	42 item OR 34 items and 1 cluster
	4	34	8 items OR 1 cluster	42 item OR 34 items and 1 cluster
	5	34	8 items OR 1 cluster	42 item OR 34 items and 1 cluster
	6	34	8 items OR 3 items and 1 cluster	42 item OR 37 items and 1 cluster
	7	34	8 items OR 1 cluster	42 item OR 34 items and 1 cluster
	8	34	8 items OR 1 cluster	42 item OR 34 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 provides the number of ELA items and Table 25 provides the percentage of mathematics items required in the blueprints by content strands, also known as subdomain or reporting category. The numbers below represent an acceptable range of items.

Table 24 provides the number of ELA items and Table 26 provides the percentage of mathematics items assessing each reporting category that appeared on the spring 2019 tests.

Table 23: Blueprint Number of Test Items Assessing Each Reporting Category in ELA

Grade	Reading Literary Text	Reading Informational Text	Listening*	Language*	Writing
3–8	14–17	14–17	0–3	3–5	1

**Note: Not reported in spring 2019*

Table 24: Observed Number of Test Items Assessing Each Reporting Category in Spring 2019 ELA

Grade	Reading Literary Text	Reading Informational Text	Listening*	Language*	Writing
3	15–16	15–16	1	4–5	1
4	14–17	15–17	1–2	4–5	1
5	15–16	14–16	1–2	5	1
6	15–16	15–16	1–2	4–5	1
7	14–16	15–16	1	5	1
8	14–16	15–16	1	5	1

**Note: Not reported in spring 2019*

Table 25: Blueprint Proportion of Test Items Assessing Each Reporting Category in Mathematics

Grade	Reporting Category	Proportion (%)
3	Operations and Algebraic Thinking	29–38
	Numbers and Operations—Base Ten and Fractions	38–47
	Measurement and Data and Geometry	24–29
4	Operations and Algebraic Thinking	24–32
	Numbers and Operations—Base Ten and Fractions	44–53
	Measurement and Data and Geometry	24–29
5	Operations and Algebraic Thinking	24–32
	Numbers and Operations—Base Ten and Fractions	41–50
	Measurement and Data and Geometry	26–32
6	Ratios and Proportional Relationships and Number System	38–47
	Expressions and Equations	29–38
	Geometry and Statistics and Probability	24
7	Ratios and Proportional Relationships and Number System	24–29
	Expressions and Equations	24–29
	Geometry	24–29
	Statistics and Probability	24–29
8	Expressions and Equations and Number System	29–38
	Functions	24–29
	Geometry and Statistics and Probability	38–47

Table 26: Observed Proportion of Test Items Assessing Each Reporting Category in Spring 2019 Mathematics

Grade	Reporting Category	Proportion (%)
3	Operations and Algebraic Thinking	29–32
	Numbers and Operations—Base Ten and Fractions	38–44
	Measurement and Data and Geometry	26–29
4	Operations and Algebraic Thinking	24–26
	Numbers and Operations—Base Ten and Fractions	44–47
	Measurement and Data & Geometry	26–29
5	Operations and Algebraic Thinking	24–26
	Numbers and Operations—Base Ten and Fractions	44–47
	Measurement and Data and Geometry	29–32
6	Ratios and Proportional Relationships and Number System	41–44
	Expressions and Equations	32–35
	Geometry and Statistics and Probability	24
7	Ratios and Proportional Relationships and Number System	24–29
	Expressions and Equations	24–26
	Geometry	24–26
	Statistics and Probability	24–26
8	Expressions and Equations and Number System	32–35
	Functions	24–26
	Geometry and Statistics and Probability	38–44

The summary tables show that the spring 2019 tests matched the blueprints at the reporting category level for both ELA and mathematics.

In addition to information about reporting categories, the 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 NH SAS provided a greater depth and breadth of learning and also fulfilled the requirements of academic rigor required by the Every Student Succeeds Act (ESSA). 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 27 shows the number of items in each DOK level in the ELA blueprint. Table 29 shows the percentage of items in each DOK level in the mathematics blueprint. Table 28 and Table 30 show the number or percentage of items in each DOK that appeared on the spring 2019 tests administered

to students. The tables show that in most cases, the number or percentage of items from each DOK level met the blueprint. Where the blueprint was not met in ELA, which occurred in DOK levels 1, 2 and 3, there was a maximum of a nine-item difference. Mathematics grade 3 had a 2% difference and grade 7 had a 1% difference in DOK level 1, a 1 % difference in grade 8 in DOK level 2, and a 1–7% difference for all grades in DOK level 3. As many of the items on the ELA Reading component were associated with passages, flexibility in testing was necessary for practical reasons.

Table 27: Blueprint Number of Items by DOK, ELA

Grades	DOK 1	DOK 2	DOK 3	DOK 4
3–8	3–10	11–18	5–13	1

Table 28: Observed Number of Items by DOK, Spring 2019 ELA

Grade	DOK 1	DOK 2	DOK 3&4	DOK 4
3	5–11	17–25	4–10	1
4	4–9	18–27	5–10	1
5	3–8	20–26	4–10	1
6	4–10	16–23	7–13	1
7	4–8	17–25	5–12	1
8	3–10	17–23	6–12	1

Table 29: Blueprint Proportion of Items by DOK, Mathematics

Grades	DOK 1	DOK 2	DOK 3
3–8	15–25%	50–65%	15–25%

Table 30: Observed Proportion of Items by DOK, Spring 2019 Mathematics

Grade	DOK 1	DOK 2	DOK 3
3	15–27%	53–64%	16–28%
4	16–24%	51–64%	17–30%
5	14–21%	51–65%	17–32%
6	15–24%	50–64%	16–32%
7	16–26%	51–65%	16–28%
8	15–24%	54–66%	16–26%

4.2 TEST CONSTRUCTION

During fall 2018, AIR psychometricians and content experts worked with NHDOE content specialists and leadership to build item pools for the spring 2019 administration. NH SAS test construction uses a structured test construction plan, explicit blueprints, and active collaborative participation from all parties. The ELA and mathematics assessments employ computer-adaptive testing that draws from item pools. For more information about AIR's adaptive algorithm, see Appendix J.

The 2019 NH SAS test item pools were built by AIR 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 discussed above, blueprints describe 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 aspect of the tests. Psychometric considerations, which ensure 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.
- Biserial correlations were reasonable and within specified bounds.
- 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 NH SAS can also be found in Volume 1.

4.3 ROLES AND RESPONSIBILITIES

4.3.1 AIR Content Team

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

- Selection of the operational items
- Revision of the operational item sets according to feedback from senior AIR content staff

- Revision of the operational item sets included according to feedback from the AIR technical team
- Revision of the operational item sets according to feedback from NHDOE
- Assistance in the generation of materials for NHDOE review
- Revision of the item pools to incorporate feedback from NHDOE

4.3.2 AIR Technical Team

The AIR 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 AIR'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 NHDOE review

4.3.3 State Content Specialists and Reviewers

NHDOE invited teachers from the field to review the proposed item pools during Content Advisory Committee and Fairness Committee meetings (see Appendices C and D, respectively, for participant information). The review process involved use of content and blueprint guidelines in addition to the statistical guidelines. NHDOE leadership was also involved in the review process for ELA and mathematics item pools 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 New Hampshire public.

NHDOE was given the opportunity to approve proposed item pools or to return them with comments to AIR's content and psychometric teams for further revision. Final approval is electronically captured in AIR's ITS and is a necessary condition for publication to our test delivery system.

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.

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

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.