

West Virginia General Summative Assessment

2021–2022

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



West Virginia DEPARTMENT OF
EDUCATION

TABLE OF CONTENTS

1.	INTRODUCTION	1
1.1	Claim Structure	1
1.2	Underlying Principles Guiding Development	2
2.	ITEM DEVELOPMENT PROCESS THAT SUPPORTS VALIDITY OF CLAIMS	3
2.1	Overview	3
2.2	Passage and Item Specifications	5
	2.2.1 Passage Specifications	5
	2.2.2 Item Specifications	8
2.3	Selection and Training of Item Writers	16
2.4	Internal Review	16
	2.4.1 Preliminary Review	17
	2.4.2 Content Review One	18
	2.4.3 Edit Review	18
	2.4.4 Senior Review	19
2.5	Review by State Personnel and Stakeholder Committees	19
	2.5.1 State Review	19
	2.5.2 Content Advisory Committee Reviews	19
	2.5.3 Language Accessibility, Bias, and Sensitivity Committee Reviews	20
	2.5.4 Markup for Translation and Accessibility Features	21
2.6	Field Testing	21
2.7	Post–Field-Test Review	21
	2.7.1 Key Verification	22
	2.7.2 Rubric Validation	22
	2.7.3 Rangefinding	23
	2.7.4 Data Review	23
3.	ICCR ITEM BANK SUMMARY	24
3.1	Current Composition of the Item Bank	24
3.2	Strategy for Pool Evaluation and Replenishment	35
4.	WVGSA TEST CONSTRUCTION	36
4.1	Test Blueprints	36
	4.1.1 ELA Blueprints	37
	4.1.2 Mathematics Blueprints	39
	4.1.3 WVGSA Test Specifications	40
4.2	Test Construction	46
	4.2.1 Paper-Based Accommodation Form Construction	47
	4.2.2 Graphical Summaries	47
4.3	Roles and Responsibilities	49
	4.3.1 CAI Content Team	49
	4.3.2 CAI Technical Team	50
	4.3.3 State Content Specialists and Reviewers	50

REFERENCES51

LIST OF TABLES

Table 1: Item Types and Descriptions, ELA25
 Table 2: Item Types and Descriptions, Mathematics.....25
 Table 3: Spring 2022 ICCR Operational and Field-Test Item Pool, ELA.....26
 Table 4: Spring 2022 ICCR Operational Item Pool, ELA26
 Table 5: Spring 2022 ICCR Field-Test Item Pool, ELA27
 Table 6: Spring 2022 ICCR Item Counts by Grade and Reporting Category, ELA.....27
 Table 7: Spring 2022 ICCR Item Counts by Grade and Depth of Knowledge (DOK), ELA28
 Table 8: Spring 2022 ICCR Item Counts by Grade and Item Type, ELA.....28
 Table 9: Spring 2022 ICCR Operational and Field-Test Item Pool, Mathematics.....30
 Table 10: Spring 2022 ICCR Operational Item Pool, Mathematics30
 Table 11: Spring 2022 ICCR Spanish Operational Item Pool, Mathematics31
 Table 12: Spring 2022 ICCR Field-Test Item Pool, Mathematics.....31
 Table 13: Spring 2022 ICCR Item Counts by Grade and Reporting Category, Mathematics.....32
 Table 14: Spring 2022 ICCR Item Counts by Grade and DOK, Mathematics.....33
 Table 15: Spring 2022 ICCR Item Counts by Item Type, Mathematics33
 Table 16: Estimated Reading Testing Times by Grade, ELA38
 Table 17: Spring 2022 Observed Testing Times by Grade, ELA.....38
 Table 18: Estimated Testing Times by Grade, Mathematics.....40
 Table 19: Spring 2022 Observed Testing Times by Grade, Mathematics.....40
 Table 20: Spring 2022 WVGSA Item Pool by Grade and Subject.....41
 Table 21: Spring 2022 Blueprint Test Length by Grade and Subject.....41
 Table 22: Spring 2022 Observed Test Length by Grade and Subject.....42
 Table 23: Blueprint Number of Test Items Assessing Each Reporting Category, ELA.....42
 Table 24: Spring 2022 Observed Number of Test Items Assessing Each Reporting Category, ELA.....43
 Table 25: Blueprint Number of Test Items Assessing Each Reporting Category, Mathematics.....43
 Table 26: Spring 2022 Observed Number of Test Items Assessing Each Reporting Category, Mathematics
44
 Table 27: Blueprint Number of Items by DOK, ELA45
 Table 28: Spring 2022 Observed Number of Items by DOK, ELA.....45
 Table 29: Blueprint Number of Items by DOK, Mathematics.....45
 Table 30: Spring 2022 Observed Number of Items by DOK, Mathematics.....45

LIST OF FIGURES

Figure 1: TCC Comparisons of Grade 4 ELA Fixed Forms48
 Figure 2: CSEM Comparison of Grade 4 ELA Fixed Forms49

LIST OF APPENDICES

- Appendix A: English Language Arts Blueprints
- Appendix B: Mathematics Blueprints
- Appendix C: Example Item Types
- Appendix D: Item Review Checklist
- Appendix E: Item Writer Training Materials
- Appendix F: Content Advisory Committee Participant Details
- Appendix G: Fairness Committee Participant Details
- Appendix H: Sample Data Review Training Materials
- Appendix I: Data Review Committee Participant Details
- Appendix J: Test Form Review Committee Participant Details
- Appendix K: ICCR Adaptive Algorithm Design

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 reflected 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 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 outlined how the items would measure each standard. The ICCR item specifications were developed in partnership with the state of Utah and began as a joint endeavor. At the outset, the ICCR item specifications matched the Student Assessment of Growth and Excellence (SAGE) item specifications. SAGE received the approval of peer reviewers, validating the quality and alignment of the specifications to the ICCR standards. Over time, the specifications have been updated to incorporate an expanding set of potential interactions and item types. An expanding pool of states has adopted the ICCR standards as a component of their item pool or, in some cases, the entire basis of their tests. The subsequent sections of this technical report will show the process that each ICCR item undergoes, including a series of stakeholder reviews in one or more participating states. While the item bank information given above remains consistent from year to year, the information in Section 3, ICCR Item Bank Summary, and Section 4, WVGSA Test Construction, is updated annually to reflect item bank growth and the observed test administration’s match to the blueprints.

1.1 CLAIM STRUCTURE

As previously stated, the assessment is designed to measure college and career readiness and to demonstrate the progress made by grades 3–11 students toward college and career readiness in mathematics and ELA.

The ELA items are designed to support the following claims about proficiency:

- 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 various 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 the following claim: *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 procedures vary by grade level and are summarized in Exhibit A.

Exhibit A: ICCR 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 ICCR 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 (TTS), translations, or assistive technologies. This goal was supported by delivering the items on Cambium Assessment, Inc.’s (CAI) Test Delivery System (TDS), 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 creating high-quality items through rigorous development processes that are managed and tracked by a content development platform which ensures that every item flows through the correct sequence of reviews and which also captures every comment and change applied to each item.

CAI sought to ensure that the items measured 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 also reviewed items for fairness and sensitivity. After the items underwent field testing, the

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

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

Combined, these principles and the supporting processes have led to an item bank that measures the standards with fidelity and does so in a manner that minimizes construct-irrelevant variance and barriers to access. The details of these processes are described further in this volume of the technical report. Organization of this Volume

This volume is organized in three sections:

1. *Overview of the Item Development Process.* This section describes the ELA and mathematics item development process that supports the validity of the claims that ICCR tests are designed to support.
2. *Overview of the ELA and Mathematics Item Pool.* This section describes the types of assessments the ELA and mathematics item pool is designed to support and methods for refreshing the pool
3. *Overview of the Test Construction Process.* This section describes the test construction for the West Virginia General Summative Assessment (WVGSA) in ELA and mathematics, including the blueprint design and test construction process.

2. ITEM DEVELOPMENT PROCESS THAT SUPPORTS VALIDITY OF CLAIMS

2.1 OVERVIEW

Cambium Assessment, Inc. (CAI) developed the Independent College and Career Readiness (ICCR) English language arts (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 all changes made to and comments associated with each item. Reviewers, including internal CAI reviewers or stakeholders in committee meetings, can review items in ITS as they will appear to the student, along with all accessibility features and tools.

The item development process begins with defining the passages and item specifications, and continues with

- selecting and training item writers;
- writing items and reviewing them internally;
- item review by state personnel and stakeholder committees;
- marking up items for translation and accessibility features;

- field testing; and
- post–field-test reviews.

Each of these steps plays a vital role in ensuring that the items can support the claims that will be based on them. Exhibit B describes how each step contributes to this goal. Each step in the development process is discussed in more detail in subsequent sections of this technical report.

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 and content limits and outlines the guidelines for meeting Depth of Knowledge (DOK) requirements and the parameters for adjusting difficulty.	Avoids using item types with accessibility constraints and provides language guidelines. Allows for multiple response modes to accommodate different styles.	
Selecting and training item writers	Ensures that item writers have the background to understand the standards and specifications. Teaches item writers how to select item types for measurement and accessibility.	Training in language accessibility, bias, and sensitivity helps item writers avoid unnecessary barriers.	
Writing items and internal reviews	Checks content and DOK alignment and evaluates and improves overall quality.	Eliminates editorial issues, and flags and removes bias and accessibility issues.	
Marking up items for translation and accessibility features		Adds universal features, such as text-to-speech (TTS) for mathematics, which reduce barriers.	Adds TTS, braille, American Sign Language (ASL), translations, and glossaries.
State personnel and stakeholder committee reviews	Checks content and DOK alignment and 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	Provides final, more focused checks on flagged items. Rubric validation and rangefinding ensure that scoring reflects standards and expectations.	Provides 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 CAI content experts. The specifications were used to develop both the Student Assessment of Growth and Excellence (SAGE) item pool and the ICCR item 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 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	1. a. The passage has a single, concrete meaning conveyed through dialogue or narration.	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).
	2. Some themes not explicitly stated	b. The main idea or theme is explicitly stated and clearly supported with supplementary details or quotes.	
	3. Multiple, successively abstract, or general levels of meaning; key theme or themes implied	c. Relationships between related concepts are clearly linked and defined.	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, Gyax, & Tapiero, 2012).
		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.	Similarly, readers draw inferences from descriptions of a character’s actions and
		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	

	<p>characters and characters’ motivations require interpretation.</p> <p>c. Mood, setting, and tone may be easily identified but do not heavily influence the overall meaning or theme.</p> <p>3.</p> <p>a. The passage contains several ideas and/or themes, both explicitly stated and implied.</p> <p>b. The reader must draw inferences about meaning from different elements of the passage, including character(s), setting, plot, dialogue, structure, and/or tone.</p> <p>c. Characters’ motivations and characteristics are strongly implied through clear action or dialogue.</p> <p>d. Mood, setting, and tone may be subtle and have a greater impact on the overall meaning.</p>	<p>stated preferences (i.e., descriptions of specific traits as being either positive or negative) (Rapp & Mensink, 2011).</p> <p>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).</p> <p>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).</p> <p>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).</p>
<p>Structure</p> <p>1. Clear, consistent narrative structure, single point of view, events in chronological order</p> <p>2. One factor varies (structure, point of view, chronology)</p> <p>3. Two or more factors vary</p> <ul style="list-style-type: none"> • Avoid requiring graphics for comprehension for accessibility reasons 	<p>1.</p> <p>a. A consistent, linear narrative is maintained throughout the passage.</p> <p>b. The narrative is presented from a single point of view and events are presented in chronological order.</p> <p>2.</p> <p>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.</p> <p>b. Changes in structure, point of view, or sequence are well-marked.</p> <p>3.</p>	<p>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).</p> <p>Flashbacks and narrator changes in a story significantly impact readers’ abilities to</p>

		<p>a. The passage contains multiple elements of complex structure, such as shifts in time, sequence, or point of view.</p> <p>b. Changes in structure, point of view, or sequence are well-marked.</p> <p>c. Elements of structure may contribute to the development of theme, setting, or plot.</p>	<p>recall or retell stories, with more flashbacks and more narrator changes throughout a story compounding this effect (Kucer, 2010).</p>
Language	<p>1. Simple, common word choice, explicit and literal use</p> <p>2. May include unfamiliar vocabulary, abstract meaning, figurative, ironic, or sarcastic use</p> <p>3. Generally dense using figurative or purposefully ambiguous, often unfamiliar language</p>	<p>1.</p> <p>a. The passage uses literal, clear, and contemporary language.</p> <p>b. High-frequency, grade-appropriate vocabulary and common word meanings are used.</p> <p>c. Syntax is simple and consistent throughout the passage.</p> <p>d. Interpretation of these words and phrases leads to a singular understanding of their role and meaning within the passage.</p> <p>2.</p> <p>a. The passage includes some unfamiliar or above-grade-level words.</p> <p>b. The meaning of most or all unfamiliar words can be determined on the basis of context clues.</p> <p>c. Familiar vocabulary may be used to convey figurative meaning.</p> <p>3.</p> <p>a. The passage includes low-frequency, domain-specific vocabulary, or uncommon word meanings.</p> <p>b. Some variation in syntax may be present.</p> <p>c. The use of figurative, ambiguous, ironic, archaic, or otherwise unfamiliar language to convey meaning is incorporated at this level.</p>	<p>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).</p> <p>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).</p> <p>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 recall and summarize a text more easily (Freebody & Anderson, 1983).</p> <p>The use of figurative language and meanings</p>

	<p>also increases the difficulty of a text. (Rommers, Dijkstra, & Bastiaansen, 2013).</p> <p>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).</p>
Total Score	<ol style="list-style-type: none"> 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 difficulties, 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 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:

1. 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.
2. 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.
3. 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. The item specifications for mathematics include the following elements:

- *Content Limits.* The content limits delineate 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.* The acceptable response mechanisms identify

the various ways in which students may respond to a prompt, such as multiple-choice, graphic response, proposition response, equation response, and multi-select items. The identified acceptable response mechanisms took accessibility concerns into consideration. For example, a graphic response item should only be used when the standard or task demand requires a graphic representation, as required when a student is required to graph 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 which practice cluster (PC) or clusters a particular standard could be aligned to: 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. It is important to note that in his recommendations on the assessment of DOK levels for mathematics, 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. He recommended that DOK 4 be assessed at the local level (Petit & Hess, 2008).
- *Task Demands.* Task demands denote the specific ways students can provide evidence of their understanding of the concept or skill. The standards are broken down into specific task demands aligned to each standard. In addition, each task demand is assigned appropriate response mechanisms, a DOK, and PCs specific to that particular task demand.
- *Relationship to Range Achievement-Level Descriptor (ALD).* Each task demand is discussed in relation to 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.* 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 a single grade 4 standard. Note that the specification provides guidance for developing items at each acceptable DOK level; it identifies the task demands and item types, and it reflects the ALDs to be included at each level. At each DOK level, the specification also provides guidance for developing items in multiple difficulty ranges.

Exhibit D: Sample Mathematics Specifications for Grade 4

Content Standard	<i>CCSS.Math.Content.4.NF</i> Number and operations — Fractions
-------------------------	---

Math.Content.4.MD.A <i>Extend understanding of fraction equivalence and ordering</i>							
<u>Math.Content.4.NF.A.2</u> 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 $\frac{1}{2}$). 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	<ul style="list-style-type: none"> *Denominators limited to 2, 3, 4, 5, 6, 8, 10, 12, 100 *Benchmarks limited to 0, $\frac{1}{4}$, $\frac{1}{2}$, $\frac{3}{4}$, 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 Multi-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.	<ul style="list-style-type: none"> • Equation response • Graphic response • Multiple-choice response • Multi-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.	<ul style="list-style-type: none"> • Multiple-choice response • Multi-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.	<ul style="list-style-type: none"> • Multi-select response • Matching response • Editing task inline response 	Students who can fluently compare various types 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.	<ul style="list-style-type: none"> • 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.	<ul style="list-style-type: none"> • Equation response • Graphic response • Multiple-choice response • Multi-select response 	Depending upon 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. <ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • 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. $1/2$ <input type="checkbox"/> $3/8$ [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.	

	<p>Select $>$, $<$, or $=$ to complete a true statement about each pair of fractions. $\frac{3}{5} \square \frac{5}{12}$ [include at least two more pairs of fractions]</p>	Medium	1, 2	Matching response	<p>This is a DOK 2 because students are comparing fractions using $<$, $>$, or $=$.</p> <p>It is medium because both fractions are less than 1.</p>
	<p>Select $>$, $<$, or $=$ to complete a true statement about each pair of fractions. $\frac{4}{3} \square \frac{6}{5}$ [include at least two more pairs of fractions]</p>	Hard	1, 2	Matching response	<p>This is a DOK 2 because students are comparing fractions using $<$, $>$, or $=$.</p> <p>It is hard because both fractions are “improper” fractions.</p>
DOK 3	<p>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.</p> <p>Part A. Generate Model Q so it is divided into 8 sections, and 5 sections are shaded.</p> <p>Then, generate Model R so it is divided into 12 sections.</p> <p>Part B. Complete the fraction comparison statement.</p> <p>Part C. Which statement is true about the two fraction models you generated and the</p>	Medium	1, 2, 3	<ul style="list-style-type: none"> • Simulation response • Editing task inline response • Multiple-choice response 	<p>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.</p> <p>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.</p>

comparison
between them?

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

- *Content Standard.* The content standard identifies the standard being assessed.
- *Content Limits.* The content limits delineate 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 the items.
- *Acceptable Response Mechanisms.* The acceptable response mechanisms identify the various ways students may respond to an item or prompt. Here, it is noted whether evidence-based selected-response (two-part items), extended-response, hot-text, multiple-choice, multi-select, and/or short-answer (to be scored automatically with our *proposition scorer*) items may be used, and if so, how.
- *DOK Demands.* The DOK demands are 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 also 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.* The sample items present a range of response mechanisms and their corresponding expected difficulties (easy, medium, and hard). Notes delineating an item’s cognitive demands and an explanation of its difficulty level are detailed for each sample item.

Exhibit E presents 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 <ul style="list-style-type: none"> Requires the student to select words or phrases from the text to answer items 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 <ul style="list-style-type: none"> Requires the student to select from four choices to answer items 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.	<ol style="list-style-type: none"> Hot-Text Response 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.	<ol style="list-style-type: none"> Hot-Text Response Multiple-Choice Response 		
DOK 3	N/A			
Item Models	Sample Item	Difficulty	Notes, Comments	Passage
DOK 1	<p>Select the sentence from the paragraph that shows why Papa had to leave the farm to go work on the railroad.</p> <p>[Hot-Text]</p>	Easy	<p>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 item 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.</p> <p>Easy Difficulty: The answer is explicitly stated in the text.</p>	<i>Roll of Thunder, Hear My Cry</i>

DOK 1	<p>Where does Brian get the idea about how to store live fish in the water?</p> <p>[Multiple-Choice]</p>	Medium	<p>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.</p> <p>Medium Difficulty: The answer is explicitly stated, but the information must be combined from details in several paragraphs.</p>	<i>Hatchet</i>
DOK 2	<p>Which sentence from the text shows that the family’s financial situation has not improved?</p> <p>[Multiple-Choice]</p>	Easy	<p>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.</p> <p>Easy Difficulty: The support for the inference stated in the item is explicitly provided in the text.</p>	<i>Roll of Thunder, Hear My Cry</i>
DOK 2	<p>Select a sentence from the text that shows that the family’s financial situation has still not improved.</p> <p>[Hot-Text]</p>	Medium	<p>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.</p> <p>Medium Difficulty: The student must choose which sentence (among all the sentences in the text) supports the inference provided in the item.</p>	<i>Roll of Thunder, Hear My Cry</i>

DOK 2	<p>Reread paragraph 6.</p> <p>Part A: Why does Papa believe the land is so important?</p> <p>Part B: Select the sentence from the text that shows why Papa thinks the land is so important.</p> <p>[two-part Hot-Text]</p>	Hard	<p>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.</p> <p>Hard Difficulty: The student must infer the answer to the item based on a character’s dialogue and then select a sentence from the text that supports this inference.</p>	<i>Roll of Thunder, Hear My Cry</i>
-------	--	------	---	-------------------------------------

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 training materials are shown in Appendix E, Item Writer Training Materials. They include

- CAI’s Language Accessibility, Bias, and Sensitivity Guidelines and
- a training module (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

ICCR’s test development structure employs 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 the items before the client review phase and provide training and feedback for all content-area team members.

ICCR items go through a rigorous, multiple-level internal review process before they are sent for 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 used by our test developers is included in Appendix D, Item Review Checklist.

The ICCR 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 requirements have been met:

- 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 item.
- 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 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 one correct option); and
- free of obvious or subtle cuing.

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. The reviewer 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 his or her own experience in test development.

2.4.3 Edit Review

During the Edit Review, editors have four primary tasks:

1. Editors perform basic line editing for correct spelling, punctuation, grammar, and mathematical and scientific notation, ensuring consistency of style across the items.
2. 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. They ensure that the answer keys are correct and that all information in the item is correct. For mathematics items, editors perform all calculations to ensure accuracy.
3. Editors review all material for fairness and language accessibility issues.
4. Editors confirm that the 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 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, both content reviewers and editors have thoroughly vetted it. Senior reviewers (in particular, senior content specialists) look at the item’s entire review history, ensuring 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 with the standard, and consistency with 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 ensure 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 before they are sent to committee review. Clients can request edits, such as wording edits, scoring edits, alignment changes, or DOK updates at this stage in the review process. 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 the requested edits will be made. At this stage, clients can either present these items to the committee (based on the edits made) or withhold them from committee review.

Wording and scoring edits cannot be made to items that have already been field tested in other states, (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 the committee.

2.5.2 Content Advisory Committee Reviews

During the Content Advisory Committee (CAC) Reviews, items are reviewed for content validity, grade-level appropriateness, and alignment with the content standards. CAC 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 (for additional information refer to Section 2.7.2, Rubric Validation).

A summary of the committee meetings appears in Exhibit F, with provides additional details about the participants, along with information regarding later meetings, in Appendix F, Content Advisory Committee Participant Details.

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
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 is presented in Exhibit G, with additional details about the participants, along with information regarding later meetings, provided in Appendix G, Fairness Committee Participant Details.

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

Location	Year	Number of Committee Members	Number of Items Reviewed	Number of Items Rejected
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 markup, whether translations or TTS, 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, items are tagged with TTS. Spanish translations, including Spanish TTS and braille, are available for a subset of items. The common ICCR Item Bank is reviewed to identify items that are appropriate for braille embossing and/or Spanish translation/Spanish TTS. The braille and translated pool include a subset of items for each grade band.

2.6 FIELD TESTING

ICCR items were embedded in operational, summative accountability assessments for field testing in participating states. CAI’s field-testing design is described in detail in Volume 1 of this report.

2.7 POST-FIELD-TEST REVIEW

After field testing, items were subjected to additional reviews that included

- key verification, for key-scored items;

- 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.

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. 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 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 using the Rubric Evaluation and Verification for Items Scored Electronically (REVISE) system. 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 illustrates the features provided by the REVISE system.

Exhibit H: Features of the REVISE Software

The image displays three screenshots of the REVISE software interface, illustrating key features:

- Sample Details:** A table showing rubric descriptions and the number of responses for different score categories.

Rubric Score Name	Rubric Description	Number of Responses
HighGridScore	Sample of responses that scored unusually high on this grid item (given overall score)	15
LowGridScore	Sample of responses that scored unusually low on this grid item (given overall score)	13
NormalResponses	Sample of responses with grid scores that are neither low nor high	17
- Responses in the sample:** A table listing individual responses and scores for a specific item.

Response	Score
18259	0
18258	0
18257	0
18256	0
18255	0
18254	0
18253	0
18252	0
18251	0
18250	0
18249	0
18248	0
18247	0
18246	0
18245	0
18244	0
18243	0
18242	0
18241	0
18240	0
18239	0
18238	0
18237	0
18236	0
18235	0
18234	0
18233	0
18232	0
18231	0
18230	0
18229	0
18228	0
18227	0
18226	0
18225	0
18224	0
18223	0
18222	0
18221	0
18220	0
18219	0
18218	0
18217	0
18216	0
18215	0
18214	0
18213	0
18212	0
18211	0
18210	0
18209	0
18208	0
18207	0
18206	0
18205	0
18204	0
18203	0
18202	0
18201	0
18200	0
18199	0
18198	0
18197	0
18196	0
18195	0
18194	0
18193	0
18192	0
18191	0
18190	0
18189	0
18188	0
18187	0
18186	0
18185	0
18184	0
18183	0
18182	0
18181	0
18180	0
18179	0
18178	0
18177	0
18176	0
18175	0
18174	0
18173	0
18172	0
18171	0
18170	0
18169	0
18168	0
18167	0
18166	0
18165	0
18164	0
18163	0
18162	0
18161	0
18160	0
18159	0
18158	0
18157	0
18156	0
18155	0
18154	0
18153	0
18152	0
18151	0
18150	0
18149	0
18148	0
18147	0
18146	0
18145	0
18144	0
18143	0
18142	0
18141	0
18140	0
18139	0
18138	0
18137	0
18136	0
18135	0
18134	0
18133	0
18132	0
18131	0
18130	0
18129	0
18128	0
18127	0
18126	0
18125	0
18124	0
18123	0
18122	0
18121	0
18120	0
18119	0
18118	0
18117	0
18116	0
18115	0
18114	0
18113	0
18112	0
18111	0
18110	0
18109	0
18108	0
18107	0
18106	0
18105	0
18104	0
18103	0
18102	0
18101	0
18100	0
18099	0
18098	0
18097	0
18096	0
18095	0
18094	0
18093	0
18092	0
18091	0
18090	0
18089	0
18088	0
18087	0
18086	0
18085	0
18084	0
18083	0
18082	0
18081	0
18080	0
18079	0
18078	0
18077	0
18076	0
18075	0
18074	0
18073	0
18072	0
18071	0
18070	0
18069	0
18068	0
18067	0
18066	0
18065	0
18064	0
18063	0
18062	0
18061	0
18060	0
18059	0
18058	0
18057	0
18056	0
18055	0
18054	0
18053	0
18052	0
18051	0
18050	0
18049	0
18048	0
18047	0
18046	0
18045	0
18044	0
18043	0
18042	0
18041	0
18040	0
18039	0
18038	0
18037	0
18036	0
18035	0
18034	0
18033	0
18032	0
18031	0
18030	0
18029	0
18028	0
18027	0
18026	0
18025	0
18024	0
18023	0
18022	0
18021	0
18020	0
18019	0
18018	0
18017	0
18016	0
18015	0
18014	0
18013	0
18012	0
18011	0
18010	0
18009	0
18008	0
18007	0
18006	0
18005	0
18004	0
18003	0
18002	0
18001	0
18000	0
- Test Item and Student Response:** A screenshot showing a math problem about plane travel with a table and a student's handwritten response.

Time (Hours)	Distance (Miles)
2	1,140
3	1,710
4	2,280

Student response: $570d$
 $1r$

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 to 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 1, Annual Technical Report, Section 4, 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 is available in Appendix H, Sample Data Review Training Materials.

Exhibit I summarizes the data review committee meetings. Details, including the composition of each committee, is available in Appendix I, Data Review Committee Participant Details.

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 Independent College and Career Readiness (ICCR) item bank is robust and has been constructed explicitly to support multiple statewide assessment programs. As described previously, ICCR items were written to the Common Core State Standards (CCSS), 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 from those that are appropriately aligned and have passed required reviews (as described in Section 2) for use on its statewide assessment. The ICCR item bank continues to grow as Cambium Assessment, Inc. (CAI) field test new items in participating states. Participating states collectively share the items and agree to field test new items yearly. 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 English language arts (ELA) and mathematics item types and provide a brief description of each. Examples of various item types can be found in Appendix C, Example Item Types.

Table 1: Item Types and Descriptions, ELA

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 several 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 several 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 several 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 CAI's Item Tracking System (ITS).

Table 2: Item Types and Descriptions, Mathematics

Response Type	Description
Equation (EQ)	Student uses a keypad with various types of mathematical symbols to create a response. Responses can include numbers, fractions, expressions, inequalities, functions, and equations.
Editing Task Choice (ETC)	Student identifies an incorrect word or phrase and chooses the replacement from several options.
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.
Multiple-Choice (MC)	Student selects one correct answer from four options.
Multiple-Select (MS)	Student selects all correct answers from several 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: The abbreviations correlate to the attributes used in CAI's Item Tracking System (ITS).

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

Table 3: Spring 2022 ICCR Operational and Field-Test Item Pool, ELA

Grade	Total Number of Items	Number of Writing Prompts
3	589	6
4	628	6
5	595	6
6	661	6
7	666	6
8	660	6
9	398	5
10	424	3
11	266	-
Total	4887	44

Table 4: Spring 2022 ICCR Operational Item Pool, ELA

Grade	Number of Total OP Items
3	500
4	537
5	506
6	575
7	577
8	567
9	371
10	360
11	266
Total	4259

Table 5: Spring 2022 ICCR Field-Test Item Pool, ELA

Grade	Number of Total Field-Test Items
3	95
4	97
5	95
6	92
7	95
8	99
9	32
10	67
Total	672

Table 6: Spring 2022 ICCR Item Counts by Grade and Reporting Category, ELA

Grade	Reading Informational Text	Reading Literary Text	Writing and Language	Speaking and Listening	Grand Total
3	210	172	111	7	500
4	203	183	144	7	537
5	193	175	127	11	506
6	256	187	117	15	575
7	234	212	121	10	577
8	247	192	193	5	567
9	145	136	83	7	371
10	164	102	91	3	360
11	128	63	71	4	266
Total	1780	1422	1058	69	4259

Table 7: Spring 2022 ICCR Item Counts by Grade and Depth of Knowledge (DOK), ELA

Grade	DOK 1	DOK 2	DOK 3	DOK 4	Grand Total
3	91	334	69	6	500
4	94	365	72	6	537
5	78	349	73	6	506
6	73	391	105	6	575
7	66	395	110	6	577
8	64	390	107	6	567
9	38	261	67	5	371
10	53	233	71	3	360
11	43	159	64	-	266
Total	600	2877	738	44	4259

Table 8: Spring 2022 ICCR Item Counts by Grade and Item Type, ELA

Grade	Item Type	Number of Items
3	Editing Task Choice	49
	Hot-Text	41
	Multiple-Choice	446
	Multiple-Select	33
	Table Match	22
	Text Entry	9
	Total	600
4	Editing Task Choice	58
	External Copy	3
	Hot-Text	45
	Multiple-Choice	448
	Multiple-Select	55
	Table Match	16
	Text Entry	11
Total	636	
5	Editing Task Choice	64
	External Copy	2
	Grid	1
	Hot-Text	54
	Multiple-Choice	404

Grade	Item Type	Number of Items
	Multiple-Select	54
	Table Match	26
	Text Entry	10
	Total	615
6	Editing Task Choice	59
	External Copy	2
	Hot-Text	40
	Multiple-Choice	503
	Multiple-Select	59
	Table Match	13
	Text Entry	11
Total	687	
7	Editing Task Choice	65
	External Copy	2
	Hot-Text	41
	Multiple Choice	464
	Multiple-Select	101
	Table Match	9
	Text Entry	12
Total	694	
8	Editing Task Choice	61
	External Copy	3
	Hot-Text	47
	Multiple-Choice	501
	Multiple-Select	55
	Table Match	11
	Text Entry	9
Total	687	
9	Editing Task Choice	57
	External Copy	1
	Grid	1
	Hot-Text	41
	Multiple-Choice	285
	Multiple-Select	29
	Table Match	2
Text Entry	7	
Total	423	
10	Editing Task Choice	64
	Hot-Text	37

Grade	Item Type	Number of Items
	Multiple-Choice	314
	Multiple-Select	34
	Table Match	3
	Text Entry	4
	Total	456
11	Editing Task Choice	41
	Hot-Text	31
	Multiple-Choice	204
	Multi-Select	27
	Table Match	1
	Text Entry	2
	Total	306
All	Grand Total	5104

Table 9: Spring 2022 ICCR Operational and Field-Test Item Pool, Mathematics

Grade	Total Number of Items
3	703
4	713
5	706
6	708
7	612
8	696
HS	1415
Total	5553

Table 10: Spring 2022 ICCR Operational Item Pool, Mathematics

Grade	Number Operational of Items
3	641
4	675
5	549

Grade	Number Operational of Items
6	688
7	473
8	552
HS	1340
Total	4918

Table 11: Spring 2022 ICCR Spanish Operational Item Pool, Mathematics

Grade	Number of Spanish OP Items
3	402
4	403
5	390
6	374
7	332
8	381
HS	961
Total	3243

Table 12: Spring 2022 ICCR Field-Test Item Pool, Mathematics

Grade	Number of Field-Test Items
3	62
4	38
5	157
6	20

Grade	Number of Field-Test Items
7	139
8	144
HS	75
Total	635

Table 13: Spring 2022 ICCR Item Counts by Grade and Reporting Category, Mathematics

Grade	Reporting Category	Number of Items
3	Geometry, Measurement and Data	186
	Number and Operations—Fractions	159
	Number and Operations in Base Ten	109
	Operations and Algebraic Thinking	187
	Total	641
4	Geometry, Measurement and Data	168
	Number and Operations—Fractions	200
	Number and Operations in Base Ten	186
	Operations and Algebraic Thinking	121
	Total	675
5	Geometry, Measurement and Data	141
	Number and Operations—Fractions	168
	Number and Operations in Base Ten	148
	Operations and Algebraic Thinking	92
	Total	549
6	Expressions and Equations	207
	Geometry	78
	Ratios and Proportional Relationships	165
	Statistics and Probability	66
	The Number System	172
	Total	688
7	Expressions and Equations	88
	Geometry	97
	Ratios and Proportional Relationships	92
	Statistics and Probability	108
	The Number System	88
	Total	473

Grade	Reporting Category	Number of Items
8	Expressions and Equations	161
	Functions	112
	Geometry	140
	Statistics and Probability	76
	The Number System	63
	Total	552
HS	Algebra	325
	Functions	358
	Geometry	410
	Number and Quantity	82
	Statistics and Probability	165
	Total	1340
All	Grand Total	4918

Table 14: Spring 2022 ICCR Item Counts by Grade and DOK, Mathematics

Grade	DOK 1	DOK 2	DOK 3	Total
3	150	405	86	641
4	156	439	80	675
5	112	362	75	549
6	177	444	67	688
7	85	308	80	473
8	127	322	103	552
HS	178	989	173	1340
Total	985	3269	664	664

Table 15: Spring 2022 ICCR Item Counts by Item Type, Mathematics

Grade	Item Type	Number of Items
3	Editing Task Choice	1
	Equation	343
	Grid	83
	Multiple-Choice	156

Grade	Item Type	Number of Items
	Multiple-Select	58
	Table Input	15
	Table Match	12
	Total	668
4	Editing Task Choice	6
	Equation	359
	Grid	56
	Multiple-Choice	118
	Multiple-Select	99
	Table Input	16
	Table Match	32
Total	686	
5	Editing Task Choice	11
	Equation	349
	Grid	29
	Multiple-Choice	161
	Multiple-Select	61
	Table Input	11
	Table Match	18
Total	640	
6	Editing Task Choice	2
	Equation	351
	Grid	43
	Multiple-Choice	199
	Multiple-Select	58
	Table Input	30
	Table Match	14
Total	697	
7	Editing Task Choice	3
	Equation	307
	Grid	37
	Multiple-Choice	155
	Multiple-Select	24
	Table Input	3
	Table Match	10
Total	539	

Grade	Item Type	Number of Items
8	Editing Task Choice	6
	Equation	252
	Grid	55
	Multiple-Choice	231
	Multiple-Select	58
	Table Input	8
	Table Match	9
	Total	619
HS	Editing Task Choice	25
	Equation	588
	Grid	66
	Hot-Text	37
	Multiple-Choice	564
	Multiple-Select	79
	Table Input	8
	Table Match	16
	Total	1383
All	Grand Total	5232

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 (EFT) slots. CAI’s field-testing design is described in detail in Volume 1, Annual Technical Report, Section 3.1.1. Currently, writing prompts are field tested in independent field tests approximately every five years.

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

1. the characteristics of the released items to be replaced,
2. the characteristics of the overused items overused in adaptive programs, and
3. the tabulations of the content coverage and ranges of difficulty that help 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 process optimizes the balance between blueprint match and individualized information. Among the many reports from the simulator are items seen by

more than 20% of students. The characteristics of these items are the primary targets for development. Overused items become candidates for release in two years once replacements have been introduced into the operational pool. For more details on the CAI item development plan, refer to Appendix L.

4. WVGSA TEST CONSTRUCTION

The West Virginia General Summative Assessment (WVGSA) in English language arts (ELA) and mathematics were constructed using items from the Independent College and Career Readiness (ICCR) bank. The tests were designed to meet the state-specific test blueprints that were written to align with the West Virginia College- and Career-Readiness Standards (WVCCRs). Because the ICCR item bank is large and contains an array of item types, the tests could be uniquely developed by drawing from its available item pool. The construction of test item pools for the online ELA and mathematics WVGSA is a process that requires expert judgment 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 further to support the claim that they are technically sound and consistent with the expectations of current professional standards.

The WVGSA is designed to support the claims described in the outset of this volume. CAI worked closely with the West Virginia Department of Education (WVDE) to create blueprints that guided the WVGSA development process. The blueprints were designed to meet the following objectives:

- To cover the full breadth and depth of the WVCCRs
- To require less than five hours of total testing time, including 60 minutes of writing
- To include machine-scored items, including true constructed-response item types, in which students must construct an equation, graph, illustration, etc.

4.1 TEST BLUEPRINTS

Test blueprints provide the parameters for the following elements:

- Test length
- The content areas to be covered and the acceptable number of items across the standards within each content area or reporting category
- The approximate number of field-test items, if applicable

The WVGSA ELA assessment includes two components, which are combined to provide overall 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

The item responses for the Writing and Reading components 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 grades 3–8 ELA 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 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 elements 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 Depth of Knowledge (DOK), included in Table 27.

Because the ICCR item bank offers a range of item types to assess all the standards described, each item pool constructed fulfills the WVGSA blueprint with various 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, testing times were estimated to be within 180 minutes for students within the 85th percentile as represented in Table 16. To estimate these Reading times, Cambium Assessment, Inc. (CAI) analyzed the average testing time for students on the 2015–2016 WVGSA. 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 and two informational passages and their associated items and language items. Specific estimates were not calculated for ELA Writing, but WVDE suggested a time allotment of two hours in the Test Administration Manual (TAM) reflected in Table 16. The observed

+ testing times in Table 17 represent the first year of test administration for the adaptive version of the WVGSA. All ELA Reading times are around or less than the estimates; the observed test times in ELA Writing fall within the recommended allotment for grades 6, 7, and 8, but exceed the allotment in grades 3, 4, and 5. The observed WVGSA testing times will be continually monitored for abnormalities over future test administrations.

Table 11: Estimated Reading Testing Times by Grade, ELA

Subject	Grade	Mean Testing Time (hours:minutes)	85th Percentile Testing Time (hours:minutes)
Reading	3	1:51	2:50
	4	1:31	2:00
	5	1:38	2:09
	6	1:33	2:04
	7	1:37	2:10
	8	1:18	1:42
Writing	3	-	2:00
	4	-	2:00
	5	-	2:00
	6	-	2:00
	7	-	2:00
	8	-	2:00

Table 12: Spring 2022 Observed Testing Times by Grade, ELA

Subject	Grade	Mean Testing Time (hours:minutes)	85th Percentile Testing Time (hours:minutes)
Reading	3	01:12	01:45
	4	01:10	01:42
	5	01:11	01:41
	6	00:58	01:20
	7	00:55	01:16
	8	00:48	01:07
Writing	3	01:23	02:13
	4	01:29	02:21
	5	01:31	02:23
	6	01:12	01:50
	7	01:10	01:45

Subject	Grade	Mean Testing Time (hours:minutes)	85th Percentile Testing Time (hours:minutes)
	8	01:03	01:36

4.1.2 Mathematics Blueprints

The blueprints developed for grades 3–8 mathematics are shown in Appendix B, Mathematics Blueprints. 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 the 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 as 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 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, CAI and WVDE 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.

Estimated testing times for mathematics, which were all expected to be well within 150 minutes, are shown in Table 18. To estimate these times, CAI 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. Using the proportion of each specific item type with regard to the item type category within the ICCR 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 2022 administration for the adaptive version of the WVGSA and are around or somewhat less than the projected times. The observed WVGSA testing times will be continually monitored for abnormalities over future test administrations.

Table 13: Estimated Testing Times by Grade, Mathematics

Grade	Mean Testing Time (hours:minutes)	85th Percentile Testing Time (hours:minutes)
G3	1:12	1:52
G4	1:16	1:56
G5	1:28	2:10
G6	1:24	2:09
G7	1:20	2:00
G8	1:07	1:49

Table 14: Spring 2022 Observed Testing Times by Grade, Mathematics

Grade	Mean Testing Time (hours:minutes)	85th Percentile Testing Time (hours:minutes)
G3	01:04	01:33
G4	01:06	01:35
G5	01:07	01:36
G6	01:01	01:24
G7	00:55	01:16
G8	00:50	01:11

4.1.3 WVGSA Test Specifications

One ELA and one mathematics item pool was constructed for each grade level 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 WVGSA 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 2022 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 test takers provide the same efforts on those items as the operational items.

Table 20 shows the number of operational and EFT items available in the WVGSA item pool during the spring 2022 test administration. Table 21 displays the blueprint requirements for operational items by grade and subject. Table 22 displays the observed number of items administered during spring 2022 for each subject and grade. Blueprint requirements were satisfied at the test level for each subject and grade.

Table 15: Spring 2022 WVGSA Item Pool by Grade and Subject

Subject	Grade	Number of Operational Items	Number of EFT Items	Total Items
Reading	3	463	118	581
	4	449	112	561
	5	445	115	560
	6	520	113	633
	7	444	115	559
	8	385	115	500
Writing	3	2	-	2
	4	2	-	2
	5	2	-	2
	6	2	-	2
	7	2	-	2
	8	2	-	2
Mathematics	3	653	83	736
	4	691	60	751
	5	561	132	693
	6	675	39	714
	7	498	105	603
	8	566	118	684

Table 16: Spring 2022 Blueprint Test Length by Grade and Subject

Subject	Grades	Number of Operational Items	Number of EFT Items or Clusters*	Total Test Length*
Reading	3–8	37–41	6–8	43–49
Writing	3–8	1	-	1
Mathematics	3–5, 7–8	34	8 items	42 items
	6	34	8 items	42 items

*Not included in the blueprints (Appendix A and Appendix B)

Table 17: Spring 2022 Observed Test Length by Grade and Subject

Subject	Grade	Number of Operational Items	Number of EFT Items	Total Test Length
Reading	3	37–40	6–8	43–48
	4	37–39	6–8	43–47
	5	37–40	6–8	43–48
	6	37–40	6–8	43–48
	7	37–41	6–8	43–49
	8	37–40	6–8	43–48
Writing	3–8	1		1
Mathematics	3	34–34	8–8	42–42
	4	34–34	8–8	42–42
	5	34–34	8–8	42–42
	6	34–34	4–4	38–38
	7	34–34	8–8	42–42
	8	34–34	8–8	42–42

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 number of mathematics items required in the blueprints by content strands, also known as subdomain or reporting category. The numbers here represent an acceptable range of items. Table 24 provides the number of ELA items and Table 26 provides the number of mathematics items assessing each reporting category that appeared on the spring 2022 tests.

Table 18: Blueprint Number of Test Items Assessing Each Reporting Category, ELA

Grade	Reading Literary Text	Reading Informational Text	Listening*	Language**	Writing**
3–5	15–17	12–14	0–3	8–10	1
6–8	12–14	15–17	0–3	8–10	1

*Not reported in spring 2022

**Reported as one category, Writing and Language

Table 19: Spring 2022 Observed Number of Test Items Assessing Each Reporting Category, ELA

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

*Not reported in spring 2022

**Reported as one category, Writing and Language

Table 20: Blueprint Number of Test Items Assessing Each Reporting Category, Mathematics

Grade	Reporting Category	Number
3	Operations and Algebraic Thinking	10–13
	Numbers and Operations—Base Ten and Fractions	13–16
	Measurement and Data and Geometry	8–10
4	Operations and Algebraic Thinking	8–11
	Numbers and Operations—Base Ten and Fractions	15–18
	Measurement and Data and Geometry	8–10
5	Operations and Algebraic Thinking	8–11
	Numbers and Operations—Base Ten and Fractions	14–17
	Measurement and Data and Geometry	9–11
6	Ratios and Proportional Relationships and Number System	13–16
	Expressions and Equations	10–13
	Geometry and Statistics and Probability	8
7	Ratios and Proportional Relationships and Number System	8–10
	Expressions and Equations	8–10
	Geometry	8–10
	Statistics and Probability	8–10
8	Expressions and Equations and Number System	10–13
	Functions	8–10
	Geometry and Statistics and Probability	13–16

Table 21: Spring 2022 Observed Number of Test Items Assessing Each Reporting Category, Mathematics

Grade	Reporting Category	Number
3	Operations and Algebraic Thinking	10–10
	Numbers and Operations—Base Ten and Fractions	14–15
	Measurement and Data and Geometry	9–10
4	Operations and Algebraic Thinking	8–9
	Numbers and Operations—Base Ten and Fractions	16–17
	Measurement and Data & Geometry	9–10
5	Operations and Algebraic Thinking	8–8
	Numbers and Operations—Base Ten and Fractions	16–16
	Measurement and Data and Geometry	10–10
6	Ratios and Proportional Relationships and Number System	14–15
	Expressions and Equations	11–12
	Geometry and Statistics and Probability	8–8
7	Ratios and Proportional Relationships and Number System	8–9
	Expressions and Equations	8–9
	Geometry	8–9
	Statistics and Probability	8–10
8	Expressions and Equations and Number System	11–12
	Functions	8–9
	Geometry and Statistics and Probability	13–14

The summary tables show that the spring 2022 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 the DOK. DOK levels are used to measure the cognitive demand of instructional objectives and assessment items. The use of DOK levels to construct the WVGSA 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 number of items in each DOK level in the mathematics blueprint. Table 28 and Table 30 show the

number of items in each DOK that appeared on the tests administered to students in spring 2022. 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 four-item difference between the blueprint and the forms. These differences occurred due to passage limits, which keep testing times down, and due to the blueprint’s need for two sets of editing tasks in the Language reporting category (6–8 items), which includes only DOK 1 items. Current item development in the ICCR bank is seeking to bolster the number and variety of DOK 3 items to mitigate this issue in the future.

Table 22: Blueprint Number of Items by DOK, ELA

Grades	DOK 1	DOK 2	DOK 3	DOK 4
3–8	6–10	15–25	6–12	1

Table 23: Spring 2022 Observed Number of Items by DOK, ELA

Grade	DOK 1	DOK 2	DOK 3	DOK 4
3	6–12	18–25	6–11	1–1
4	5–10	18–25	6–11	1–1
5	5–10	19–25	6–12	1–1
6	6–14	17–24	6–12	1–1
7	5–11	17–26	6–11	1–1
8	5–11	18–25	6–12	1–1

Table 24: Blueprint Number of Items by DOK, Mathematics

Grade	DOK 1	DOK 2	DOK 3
3	5–9	17–22	5–9
4	5–9	17–22	5–9
5	5–9	17–22	5–9
6	5–9	17–22	5–9
7	5–9	17–22	5–9
8	5–9	17–22	5–9

Table 25: Spring 2022 Observed Number of Items by DOK, Mathematics

Grade	DOK 1	DOK 2	DOK 3
3	6–8	19–21	6–8

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

4.2 TEST CONSTRUCTION

During fall 2021 CAI psychometricians and content experts worked with WVDE content specialists and leadership to build item pools for the spring 2022 administration. WVGSA test construction utilizes 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 CAI’s adaptive algorithm refer to Appendix K, ICCR Adaptive Algorithm Design.

CAI test developers built the 2022 WVGSA test item pools to match items exactly to the detailed test blueprints and target item difficulty and test information distribution. Operational items were selected to fulfill the blueprint for each grade. The subsequent sections of this technical report 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 previously, blueprints describe the content to be covered, the DOK with which it will be covered, the item types that will measure the constructs, and other content-relevant aspect of the tests. The psychometric considerations that ensure that students receive scores with similar precision, include ensuring the following:

- A reasonable range of item difficulties was included.
- The p -values for the items were reasonable and within specified bounds.
- The biserial correlations were reasonable and within specified bounds.
- The item response theory (IRT) a -parameters were reasonable and greater than 0.40 for all items.
- The IRT b -parameters were reasonable for all items.
- The IRT c -parameters were less than 0.40 for multiple-choice items.

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

4.2.1 Paper-Based Accommodation Form Construction

Student scores should not depend upon the mode of administration or type of test form. The braille tests are the only paper-based, fixed form accommodated tests for WVGSA. Of note, scores obtained via alternate modes of administration must be established as comparable to scores obtained through online testing. This section outlines the overall test development plans that ensured comparability between the online and paper-based tests.

To build paper-based braille forms, content specialists began with the online pool and removed any items that they could not render on paper and would be inaccessible to visually impaired students taking the braille tests. Next, content specialists constructed fixed-forms adhering to the test blueprint. All overall, reporting category, DCI, and performance standard level blueprint requirements were met.

4.2.2 Graphical Summaries

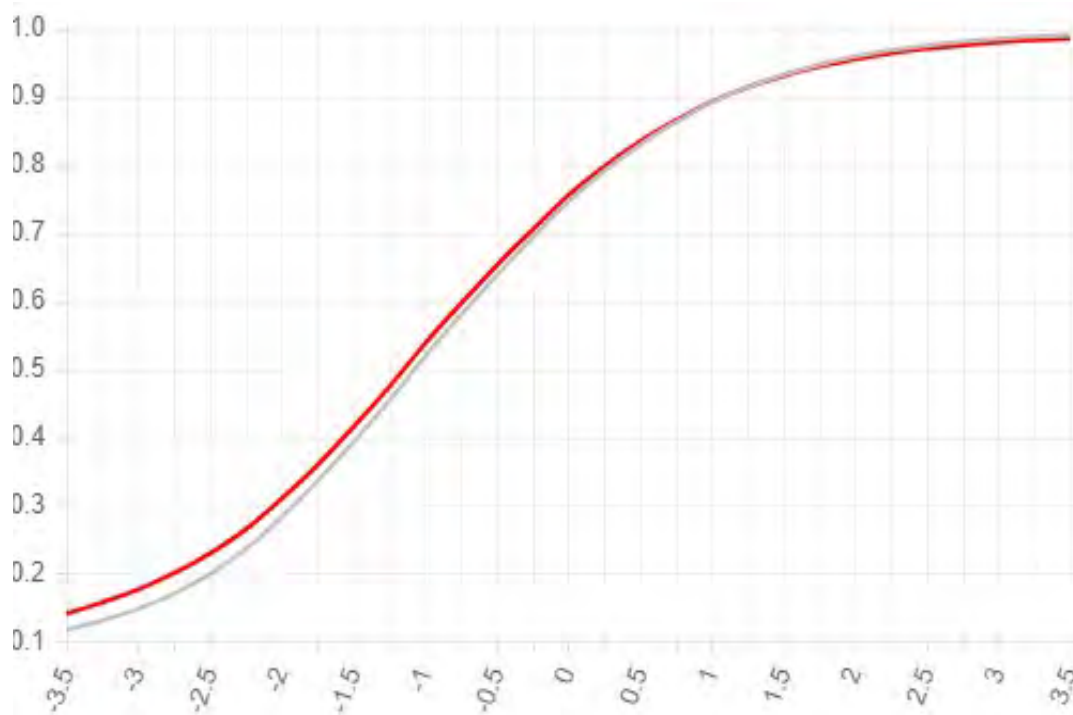
In the construction of paper-based forms, psychometricians and content specialists use graphical summaries for visualization in addition to comparing item statistics between the two forms.

Test Characteristic Curve

An item characteristic curve (ICC) shows the probability of a correct response as a function of ability, given an item's parameters. Test characteristic curves (TCCs) can be constructed as the sum of ICCs for the items included on any given test. The TCC can be used to determine test taker raw scores or percentage-correct scores that are expected at a given ability level. When two tests are developed to measure the same ability, their scores can be equated using TCCs.

The spring 2021 DEI paper-based form TCCs were the targets for the spring 2022 forms. Items were selected for paper-based such that the 2022 form TCCs matched the 2021 form TCCs as closely as possible. Figure 1 compares the TCCs for both base and newly constructed grade 4 ELA forms.

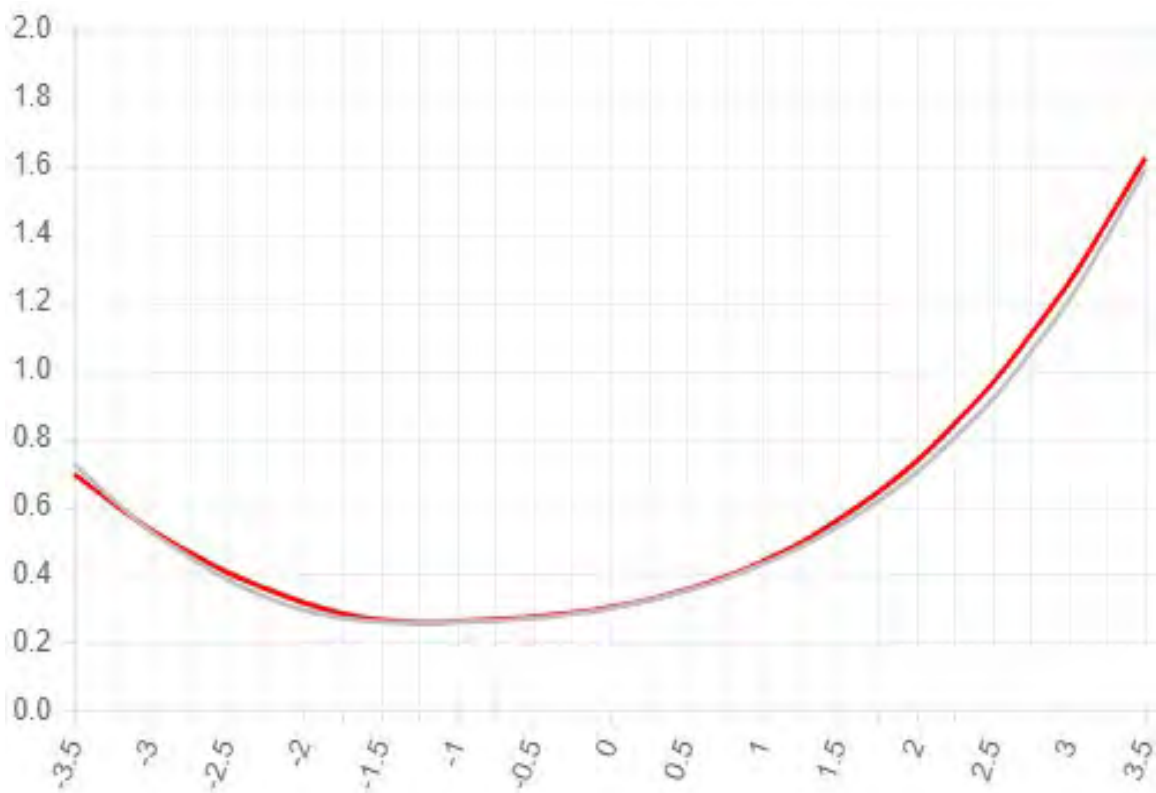
Figure 1: TCC Comparisons of Grade 4 ELA Fixed Forms



Conditional Standard Error of Measurement Curve

The Conditional Standard Error of Measurement (CSEM) curve shows the level of error of measurement expected at each ability level. The CSEM is calculated as the reciprocal of the square root of the test information function, and thus the CSEM is lowest when information is highest. Ability estimates in the middle of the distribution often appear more reliable than the ability estimates at the high and low ends of the scale. Figure 2 compares the CSEMs of both base and newly constructed grade 4 ELA forms.

Figure 2: CSEM Comparison of Grade 4 ELA Fixed Forms



4.3 ROLES AND RESPONSIBILITIES

4.3.1 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:

- Selecting the operational items
- Revising the operational item sets according to feedback from senior CAI content staff
- Revising the operational item sets according to feedback from CAI psychometric staff
- Revising the operational item sets according to feedback from WVDE
- Assisting in the generation of materials for WVDE review
- Revising the item pools to incorporate feedback from WVDE

4.3.2 CAI Technical Team

The CAI technical team, which includes psychometricians and statistical support associates, prepares the item bank by updating the Item Tracking System (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 WVDE review

4.3.3 State Content Specialists and Reviewers

WVDE invited teachers from the field to review the proposed item pools during Content Advisory Committee and Fairness Committee meetings (refer to Appendix F, Content Advisory Committee Participant Details and Appendix G, Fairness Committee Participant Details, respectively, for participant information). The review process involved use of the content and blueprint guidelines in addition to the statistical guidelines. WVDE 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 test takers, and the acceptability of test content to the West Virginia public.

WVDE was given the opportunity to approve proposed item pools or to return them with comments to CAI’s content and psychometric teams for further revision. Final approval is electronically captured in CAI’s ITS and is a necessary condition for publication to our TDS.

REFERENCES

- 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.
- 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: 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), 21–27.
- 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.

Appendix A
English Language Arts Blueprints

English Language Arts Blueprint

WVGSA Grades 3-5



WVGSA ELA Blueprint - Grades 3-5		
Domain	Number of Items	
Cluster	Min	Max
Reading Literary Text	15**	17**
• Key Ideas and Details	6	8
• Craft and Structure	6	8
• Integration of Knowledge and Ideas	1	3
Reading Informational Text	12**	14**
• Key Ideas and Details	5	7
• Craft and Structure	4	6
• Integration of Knowledge and Ideas	1	3
• Listening	0	3
Writing and Language	*	*
• Writing	1	1
• Language	8	10

Reading Passages	Min	Max
• Literary passages	2	2**
• Literary paired passages	1	2
• Informational passages	2	2**
• Informational paired passages	1	2

DOK Ranges	Min	Max
DOK 1	6	10
DOK 2	15	25
DOK 3	6	12
DOK 4 (Writing)	1	1

* There is no minimum and maximum number of items listed for Writing and Language. There will be one written student response and 8-10 language items.

**Indicates a reduction in items and/or passages from 2017-2018 WVGSA Blueprint.

English Language Arts Blueprint

WVGA Grades 6-8



WVGA ELA Blueprint - Grades 6-8		
Domain	Number of Items	
Cluster	Min	Max
Reading Literary Text	12**	14**
• Key Ideas and Details	5	7
• Craft and Structure	4	6
• Integration of Knowledge and Ideas	1	3
Reading Informational Text	15**	17**
• Key Ideas and Details	6	8
• Craft and Structure	6	8
• Integration of Knowledge and Ideas	1	3
• Listening	0	3
Writing and Language	*	*
• Writing	1	1
• Language	8	10

Reading Passages	Min	Max
• Literary passages	2	2**
• Literary paired passages	1	2
• Informational passages	2	2**
• Informational paired passages	1	2

DOK Ranges	Min	Max
DOK 1	6	10
DOK 2	15	25
DOK 3	6	15
DOK 4 (Writing)	1	1

* There is no minimum and maximum number of items listed for Writing and Language. There will be one written student response and 8-10 language items.

**Indicates a reduction in items and/or passages from 2017-2018 WVGA Blueprint.

Appendix B
Mathematics Blueprints

Mathematics Blueprint

WVGSA Grade 3



WVGSA Blueprint - Grade 3			
Domain		Number of Items	
Cluster		Min	Max
OA	Operations and Algebraic Thinking	13	16
	Represent and solve problems involving multiplication and division	1	6
	Understand the properties of multiplication and the relationship between multiplication and division	1	5
	Multiply and divide within 100	0	2
	Solve problems involving the four operations, and identify and explain patterns in arithmetic	1	5
NBT & NF	Number and Operations in Base Ten & Fractions	16	19
	Use place value understanding and properties of operations to perform multi-digit arithmetic	8	11
	Develop understanding of fractions as numbers	8	11
MD & G	Measurement, Data and Geometry	10	12
	Solve problems involving measurement and estimation of intervals of time, liquid volumes, and masses of objects	1	4
	Represent and interpret data	1	4
	Geometric measurement: understand concepts of area and relate area to multiplication and division	1	4
	Geometric measurement: recognize perimeter as an attribute of plane figures and distinguish between linear and area measures	0	2
	Reason with shapes and their attributes	1	4
DOK Ranges		Min	Max
DOK 1		7	11
DOK 2		22	27
DOK 3		6	11
Mathematical Habits of Mind Sub-Score		Min	Max
MHM: Modeling and Problem Solving		10	14
MHM: Use Mathematical Reasoning		10	14

Mathematics Blueprint

WVGSA Grade 4



WVGSA Blueprint - Grade 4			
Domain		Number of Items	
Cluster		Min	Max
OA	Operations and Algebraic Thinking	10	13
	Use the four operations with whole numbers to solve problems	3	7
	Gain familiarity with factors and multiples	1	4
	Generate and analyze patterns	1	4
NBT & NF	Number and Operations in Base Ten & Fractions	19	22
	Generalize place value understanding for multi-digit whole numbers	3	6
	Use place value understanding and properties of operations to perform multi-digit arithmetic	3	6
	Extend understanding of fraction equivalence and ordering	1	4
	Build fractions from unit fractions by applying and extending previous understandings of operations on whole numbers	1	4
Understand decimal notation for fractions, and compare decimal fractions	3	6	
MD & G	Measurement, Data and Geometry	10	12
	Solve problems involving measurement and conversion of measurements from a larger unit to a smaller unit	1	5
	Represent and interpret data	0	2
	Geometric measurement: understand concepts of angle and angle measure	1	6
Draw and identify lines and angles, and classify shapes by properties of their lines and angles	3	6	
DOK Ranges		Min	Max
DOK 1		7	11
DOK 2		22	27
DOK 3		6	11
Mathematical Habits of Mind Sub-Score		Min	Max
MHM: Modeling and Problem Solving		8	12
MHM: Use Mathematical Reasoning		8	12

Mathematics Blueprint

WVGSA Grade 5



WVGSA Blueprint - Grade 5			
Domain		Number of Items	
Cluster		Min	Max
OA	Operations and Algebraic Thinking	10	14
	Write and interpret numerical expressions	2	10
	Analyze patterns and relationships	1	5
NBT & NF	Number and Operations in Base Ten & Fractions	17	21
	Understand the place value system	1	6
	Perform operations with multi-digit whole numbers and with decimals to hundredths	1	6
	Use equivalent fractions as a strategy to add and subtract fractions	1	5
	Apply and extend previous understandings of multiplication and division to multiply and divide fractions	1	7
MD & G	Measurement, Data and Geometry	11	14
	Convert like measurement units within a given measurement system	0	2
	Represent and interpret data	0	2
	Geometric measurement: understand concepts of volume and relate volume to multiplication and to addition	1	6
	Graph points on the coordinate plane to solve real-world and mathematical problems	0	2
	Classify two-dimensional figures into categories based on their properties	0	2
DOK Ranges		Min	Max
DOK 1		7	11
DOK 2		22	27
DOK 3		6	11
Mathematical Habits of Mind Sub-Score		Min	Max
MHM: Modeling and Problem Solving		8	12
MHM: Use Mathematical Reasoning		8	12

Mathematics Blueprint

WVGSA Grade 6



WVGSA Blueprint - Grade 6			
Domain		Number of Items	
Cluster		Min	Max
Non-Calculator Total			
RP & NS	Ratios and Proportional Relationships & Number System	16	19
	Understand ratio concepts and use ration reasoning to solve problems	5	10
	Apply and extend previous understandings of multiplication and division to divide fractions by fractions	0	2
	Compute fluently with multi-digit numbers and find common factors and multiples	1	5
	Apply and extend previous understandings of numbers to the system of rational numbers	1	5
EE	Expressions and Equations	13	16
	Apply and extend previous understandings of arithmetic to algebraic expressions	3	9
	Reason about and solve one-variable equations and inequalities	3	9
	Represent and analyze quantitative relationships between dependent and independent variables	0	2
Calculator Total			
G & SP	Geometry & Statistics and Probability	10	10
	Solve real-world and mathematical problems involving area, surface area, and volume	1	7
	Develop understanding of statistical variability	1	5
	Summarize and describe distributions	1	4
DOK Ranges		Min	Max
DOK 1		7	11
DOK 2		22	27
DOK 3		6	11
Mathematical Habits of Mind Sub-Score		Min	Max
MHM: Modeling and Problem Solving		10	14
MHM: Use Mathematical Reasoning		8	10

Mathematics Blueprint

WVGSA Grade 7



WVGSA Blueprint - Grade 7			
Domain		Number of Items	
Cluster		Min	Max
Calculator Total			
RP & NS	Ratios and Proportional Relationships & Number System	10	12
	Analyze proportional relationships and use them to solve real-world and mathematical problems	3	7
	Apply and extend previous understandings of operations with fractions to add, subtract, multiply, and divide rational numbers	3	7
EE	Expressions and Equations	10	12
	Use properties of operations to generate equivalent expressions	3	7
	Solve real-life and mathematical problems using numerical and algebraic expressions and equations	3	7
G	Geometry	10	12
	Draw, construct, and describe geometrical figures and describe the relationships between them	3	7
	Solve real-life and mathematical problems involving angle measure, area, surface area, and volume	3	7
SP	Statistics and Probability	10	12
	Use random sampling to draw inferences about a population	1	4
	Draw informal comparative inferences about two populations	1	4
	Investigate chance processes and develop, use, and evaluate probability models	3	7
DOK Ranges		Min	Max
DOK 1		7	11
DOK 2		22	27
DOK 3		6	11
Mathematical Habits of Mind Sub-Score		Min	Max
MHM: Modeling and Problem Solving		10	14
MHM: Use Mathematical Reasoning		9	22

Mathematics Blueprint

WVGSA Grade 8



WVGSA Blueprint - Grade 8			
Domain		Number of Items	
Cluster		Min	Max
Calculator Total			
EE & NS	Expressions and Equations & Number System	12	16
	Know that there are numbers that are not rational, and approximate them by rational numbers	0	2
	Work with radicals and integer exponents	1	5
	Understand connections between proportional relationships, lines, and linear equations	1	5
	Analyze and solve linear equations and pairs of simultaneous linear equations	1	5
F	Functions	10	12
	Define, evaluate, and compare functions	3	7
	Use functions to model relationships between quantities	3	7
G & SP	Geometry & Statistics and Probability	16	20
	Understand congruence and similarity using physical models, transparencies, or geometry software	2	7
	Understand and apply the Pythagorean Theorem	1	5
	Solve real-world problems involving volume cylinders, cones, and spheres.	0	2
	Investigate patterns of association in bivariate data	2	7
DOK Ranges		Min	Max
DOK 1		7	11
DOK 2		22	27
DOK 3		6	11
Mathematical Habits of Mind Sub-Score		Min	Max
MHM: Modeling and Problem Solving		9	12
MHM: Use Mathematical Reasoning		8	11

Appendix C
Example Item Types

Item Types Available in the West Virginia Assessments

Selected-Response Item Types

Multiple-Choice Interactions

Multiple-choice (MC) interactions require students to select a single option from a list of possible answer options. The number and orientation of answer options in a multiple-choice interaction are configurable. Answer options may appear vertically, horizontally, vertically stacked (in a specified number of columns), or horizontally stacked (in a specified number of rows).

What is the product of 68 and 90?

A 612

B 1,260

C 6,120

D 6,300

Multiple-Select (MS) Interactions

Multiple-select interactions require students to select one or more options from a list of possible answer options. The number and orientation of answer options in a multiple-select interaction are configurable. Answer options may appear vertically, horizontally, horizontally stacked (in a specified number of rows), or vertically stacked (in a specified number of columns). In the example which follows, the options are stacked.

Select the values that are greater than or equal to $\frac{1}{2}$.

0.6 .45

$\frac{2}{6}$ One Fifth

$\frac{5}{8}$ $\frac{2}{10}$

Evidence-Based Selected-Response Interactions (ELA only)

Evidence-based selected-response (EBSR) interactions include two parts, Part A and Part B. In Part A, students respond to a multiple-choice question with only one answer. In Part B, students are presented with options that are designed to support their answer in Part A. These options can either be in multiple choice (one correct answer) or multiple select (multiple correct answers) formats.

Part A

What does the first story show about Iggy and Sal?

- Ⓐ Iggy is a better reader than Sal is.
- Ⓑ Sal imagines things more than Iggy.
- Ⓒ Sal and Iggy have only been partners for a short time.
- Ⓓ Sal and Iggy try to be better than the other at solving cases.

Part B

Which statement from the story supports the response in Part A?

- Ⓐ “I settled into my recliner with *The Insect Informer* and scanned the headlines.”
- Ⓑ “Sal bounded into the room. ‘Get this, Ig. Our new client?’”
- Ⓒ “‘And she says an *alien* did it!’”
- Ⓓ “I smiled, excited that I was going to crack this case before my big-brained pal Sal.”

Table Match Interactions

Table match (MI) interactions arrange two sets of match options in a table, with one set listed in columns and the other set listed in rows. Students match options in the columns to options in the rows by marking checkboxes in the cells where the columns and rows intersect.

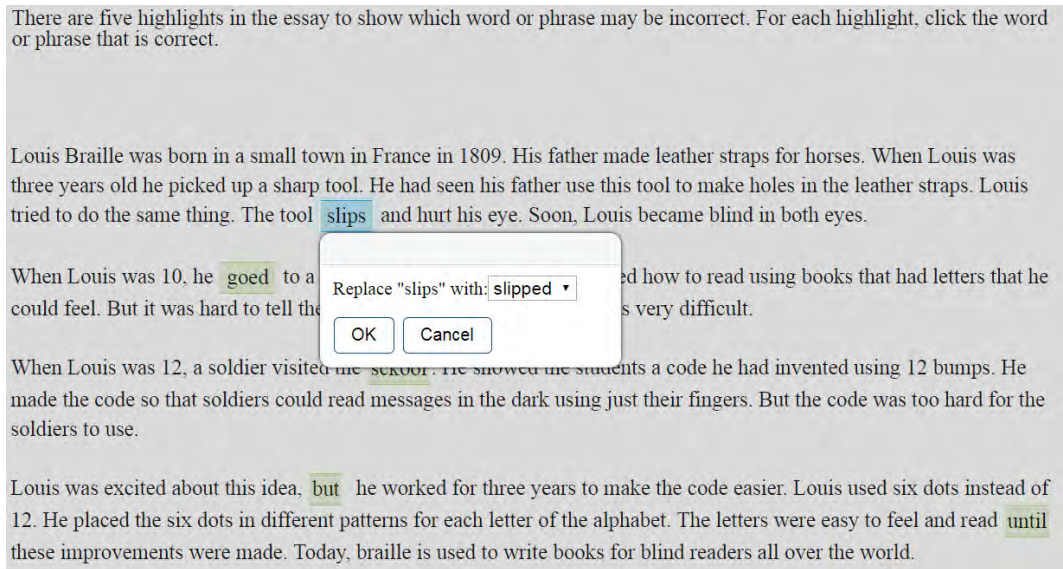
For each number listed in the rows of the table, mark the checkboxes for each column that describes that number.

	Perfect Square	Prime Number	Odd Number	Even Number
5	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
12	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Table match interactions allow the test developer to customize the number of match options in each set and enter the content for each match option. The test developer can also set restrictions on the number of matches students can make.

Edit Task with Choice Interactions (ELA only)

Edit task with choice (ETC) interactions provide students with a sentence or paragraph containing one or more tagged text elements. Tagged elements usually contain an error, such as improper spelling or grammar. To respond to these interactions, students click a tagged element to replace the tagged text elements with options selected from a drop-down list. The entered text replaces the original tagged text.



Edit task interactions allow the test developer to enter the text that appears in the response area and tag elements within the text that students can edit.

Hot Text Interactions

The Hot Text Interaction Editor allows the test developer to create content for the following interaction types:

- **Error! Reference source not found.**
- **Error! Reference source not found.**
- **Error! Reference source not found.**

Selectable Hot Text Interactions

Selectable hot text (HT) interactions require students to select one or more text elements in the response area.

Select the sentences that support the inference that the area is in danger of losing its moose population. Select **all** that apply.

A similar boom-and-bust cycle occurs between predator and prey. Ten times the size of a wolf, a moose has long, strong legs and a dangerous kick. So wolves prey mainly on old and weak animals. Good hunting means food for the whole pack. Wolves then raise lots of pups, and their numbers increase. **More wolves mean more mouths to feed and more moose get eaten.** However, when the moose population decreases, wolves starve.

Selectable hot text interactions allow the test developer to set the minimum and maximum number of elements students can select, enter the text that appears in the response area, and tag the text elements that will be selectable.

Re-orderable Hot Text Interactions

Re-orderable hot text (HT) interactions require students to click and drag hot text elements into a different order.

Place the following sentences in the correct order.

Hey Jude. And make it better. Don't be afraid. Take a sad song.

Re-orderable hot text interactions allow the test developer to enter the re-orderable text elements in the response area. The test developer can specify the elements' orientation and set them to appear in random order to students.

Drag-from-Palette Hot Text Interactions (a.k.a. Hot Text Gap Match)

Drag-from-palette hot text (HT) interactions require students to drag elements from a palette into the available blank table cells or "gaps" (text boxes) in the response area. Palette elements may consist of

text and/or images. Students may be able to drag the same palette element into multiple gaps, depending on the interaction's configuration.

Drag and drop the characteristics into the appropriate table cells below.

Fortunato's character	Montessor's character

Sinister and calculating
Cowardly and irreverent
Egotistical and rude
Lazy and inconsiderate

Drag-from-palette hot text interactions allow the test developer to enter the elements that appear in the palette, enter static text for the response area, and create the gap targets where students can drag the text elements. The test developer can enter all of the elements in a single text box or enter each segment in its own text box.

- Can set a minimum/maximum number of times a student is required/allowed to use a specific palette object
- Only supports drag-and-drop of palette items (images or plain text) onto pre-defined drop targets (“gaps” or “blanks”) in the body text
 - These palette items are always confined to a special palette region (no “preplacing” them).
 - There is some control over palette placement.
 - The items can be placed only in predefined “target” regions.

Machine-Scored Constructed-Response Item Types

Table Input Interactions (Mathematics Only)

Table input (TI) interactions provide students with a table that includes one or more blank cells. Each blank cell displays a text box in which students can type their numeric response.

The total number of hours, t , that Trent has practiced his guitar after d days is modeled by the equation shown.

$$t = 3d$$

Complete the table to describe this relationship.

d	t
1	<input type="text"/>
3	<input type="text"/>
<input type="text"/>	15

Table input interactions allow the test developer to customize the number of rows and columns in the table, specify which cells display text boxes, and enter content for the read-only cells.

Extended-Response Interactions (ELA only)

Extended-response (ER) interactions require students to type a response in a text box. Extended-response interactions are scored by an uploaded essay scoring model that analyzes the student's response to identify variations of acceptable key words and phrases. For extended-response interactions, the test developer can allow the test developer to specify the maximum response length for the text box and the type of text editor available to students.

Select a sentence in the passage that does not fit with the overall structure and explain why it is disruptive to the organization of the passage.

Type your answer in the space provided.

Equation Interaction Editor (Mathematics Only)

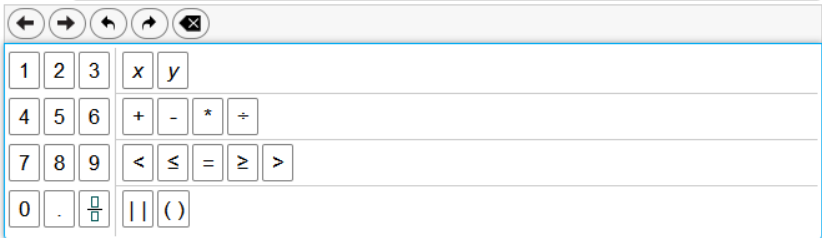
The Equation Interaction Editor allows the test developer to create content for equation (EQ) interactions only. Equation interactions require students to enter a response into input boxes using an on-screen keypad, which may consist of special mathematics characters. Students can also enter their

response via a physical keyboard, but they cannot enter any characters that are not included in the on-screen keyboard.

Use the quadratic formula to find the values of x for the following equation:
 $y = x^2 + 2x - 3$

X =

X =



The keypad interface includes navigation buttons (back, forward, undo, redo, clear) and a grid of mathematical symbols and numbers:

1	2	3	x	y			
4	5	6	+	-	*	÷	
7	8	9	<	≤	=	≥	>
0	.	$\frac{\square}{\square}$			()		

Equation interactions allow the test developer to select the buttons to include in the on-screen keypad, enter static text in the response area, and specify the number of input boxes to include in the response area. When selecting buttons to include in the keypad, the test developer can add individual buttons or an entire row or tab of buttons.

Student responses can include any of the following:

- Numeric values only (e.g., integers, decimals, fractions/mixed numbers)
- Expressions
- Inequalities
- Functions
- Equations

Grid Interactions Types (Mathematics Only)

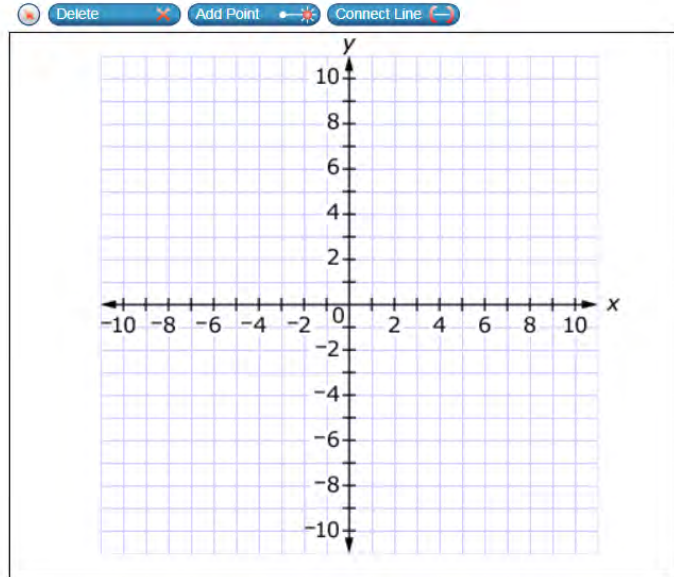
Grid (GI) interactions require students to enter a response by interacting with a grid area in the answer space. There are three general ways in which students can interact with the grid area.

- **Graphing Functionality:** Students can use various tool buttons to add points, lines, and other geometric shapes to the grid area. Only the grid interaction sub-type allows the test developer to create interactions with this functionality.

10



Use the Connect Line tool to create a rectangle with an area of 35 square units and one side with vertices at (1, 3) and (1, -4).



- **Hot Spot Functionality:** Students can click or hover over interactive regions in the grid area (hot spots) in order to activate them. Activated hot spots become highlighted, become outlined, or display an image.

7

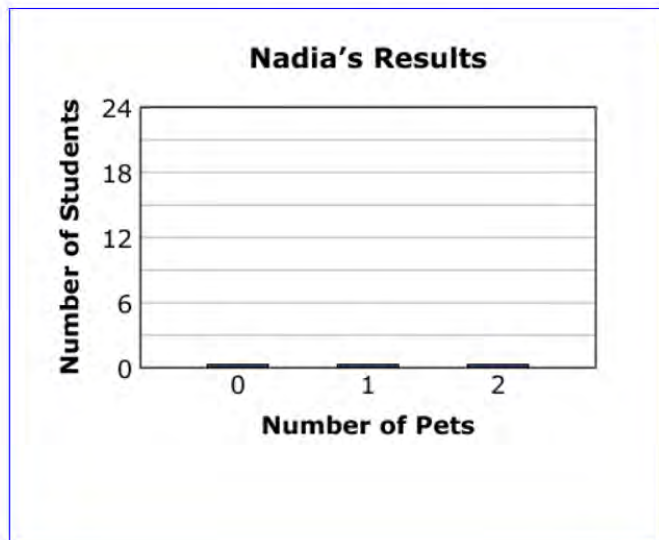


Nadia asks each student in her class how many pets he or she has. The results are shown in the table.

Nadia's Results

Number of Pets	Number of Students
0	15
1	18
2	6

Click between the lines to create a bar graph that shows Nadia's results.



- **Drag-and-Drop Functionality:** Students can click image or text objects and drag them into various locations in the grid area. The objects for these interactions are either provided in a palette beside the grid area or pre-placed within the grid area itself.

- These palette items can be “preplaced” on the canvas or listed in a separate palette.
- The items can be placed anywhere on the canvas or guided to specific regions with snap points.



Appendix D
Item Review Checklist

Item Review Checklist

1. Passage Set: Is the item unique? Does it assess new material not already tested in the passage set? Is the item free of cueing and clanging with other questions in the set?
2. Does the item fit the standard/target? Is there a more appropriate standard/target?
3. Is the content accurate and grade appropriate (use vocabulary resources when appropriate)?
4. If the item is based on a passage, is it passage-based, including all distractors?
5. Is the language clear, concise, consistent, and appropriate for the grade level? Does each word serve a purpose?
6. Does the stem or prompt flow well, if not seamlessly, to the options or task?
7. Are the options parallel and plausible? Can you justify them with sufficient rationales? Are the options free of echo, or word(s), repeated from the stem?
8. Is the item fair, accessible, and free from bias and sensitivity issues?
9. Does the item test a single construct? Keep in mind, the more difficult the content/task/construct, the more precise the language should be.
10. Is there anything confusing about the item (for example, options that are subsets of the key- or non-passage-based options, or options that are true in one sense, but not true as they apply to the stem)?
11. Is the Depth of Knowledge/complexity accurate and appropriate?
12. Does the item assess significant information?
13. Is the item type appropriate? Does the content lend itself to, if not require, this format?

Strategies for Editing Text to Produce Plain Language

- Reduce excessive length.
- Use common words.
- Avoid ambiguous words.
- Limit irregularly spelled words.
- Avoid inconsistent naming and graphic conventions.
- Avoid multiple terms for the same concept.
- Limit the use of embedded clauses and phrases.
- Avoid the passive voice.

Appendix E
Item Writer Training Materials

Exhibit 1: LABS Guidelines



LABS GUIDELINES

1 STEREOTYPING

Testing materials should not present persons stereotyped according to the following characteristics:

- Age
- Disability
- Gender
- Race/Ethnicity
- Sexual orientation

2 SENSITIVE OR CONTROVERSIAL SUBJECTS

Controversial or potentially distressing subjects should be avoided or treated sensitively. For example, a passage discussing the historical importance of a battle is acceptable whereas a graphic description of a battle would not be. Controversial subjects include:

- | | |
|-----------------------|-----------------|
| a. Death and Disease | e. Religion |
| b. Gambling* | f. Sexuality |
| c. Politics (Current) | g. Superstition |
| d. Race relations | h. War |

**References to gambling should be avoided in mathematics items related to probability.*

3 ADVICE

Testing materials should not advocate specific lifestyles or behaviors except in the most general or universally agreed-upon ways. For example, a recipe for a healthful fruit snack is acceptable but a passage recommending a specific diet is not. The following categories of advice should be avoided:

- Religion
- Sexual preference
- Exercise
- Diet

4 DANGEROUS ACTIVITY

Tests should not contain content that portrays people engaged in or explains how to engage in dangerous activities. Examples of dangerous activities include:

- Deep-sea diving

- Stunts
- Parachuting
- Smoking
- Drinking

5 POPULATION DIVERSITY AND ETHNOCENTRISM

Testing materials should:

- Reflect the diversity of the testing population
- Use stimulus materials (such as works of literature) produced by members of minority communities
- Use personal names from different ethnic origin communities
- Use pictures of people from different ethnic origin communities
- Avoid ethnocentrism (the attitude that all people should share a particular group’s language, beliefs, culture, or religion)

6 DIFFERENTIAL FAMILIARITY AND ELITISM

Specialized concepts and terminology extraneous to the core content of test questions should be avoided. This caveat applies to terminology from the fields of:

- Construction
- Finance
- Sports
- Law
- Machinery
- Military topics
- Politics
- Science
- Technology
- Agriculture

7 LANGUAGE USE

Language should be as inclusive as possible.

- Avoid “Man” words like mankind, manmade, and the generic “he”
- Use equal pairs such as husband and wife rather than man and wife

8 LANGUAGE ACCESSIBILITY

The grammar and vocabulary should be clear, concise, and appropriate for the intended grade level. The following should be avoided or used with care:

- Passive constructions
- Idioms
- Multiple subordinate clauses
- Pronouns with unclear antecedents
- Multiple-meaning words

- Nonstandard grammar
- Dialect
- Jargon

9 ILLUSTRATIONS AND GRAPHICS

Illustrations and graphics should embody all of the previously referenced LABS Guidelines.

Exhibit 2: LABS Checklist



LABS – Checklist

Stereotyping Considerations

- Does the material negatively represent or stereotype people based on gender or sexual preference?
- Does the material portray one or more people with disabilities in a negative or stereotypical manner?
- Does the material portray one or more religious groups as aggressive or violent?
- Does the material romanticize or demean people based on socioeconomic status?
- Does the material portray one or more ethnic groups or cultures participating in certain stereotypical activities or occupations?
- Does the material portray one or more age groups in a negative or stereotypical manner?

Sensitive / Controversial Material Considerations

- Does the material require a student to take a position that challenges authority?
- Does the material present war or violence in an overly graphic manner?
- Does the material present sensitive or highly controversial subjects, such as death, war, abortion, euthanasia, or natural disasters, except where they are needed to meet State Content Standards?
- Does the material require examinees to disclose values that they would rather hold confidential?
- Does the material present sexual innuendoes?
- Does the material trivialize significant or tragic human experiences?
- Does the material require the parent, teacher, or examinee to support a position that is contrary to their religious beliefs?

Advice Considerations

- Does the material contain advice pertaining to health and well-being about which there is not universal agreement?

Population Diversity

- Is the material written by members of diverse groups?
- Does the material reflect the experiences of diverse groups?
- Does the material portray people in positive nontraditional roles?

- Does test material represent the racial and ethnic composition of the testing population?
- Does the material reflect ethnocentrism?
- Does the material refer to population subgroups accurately?
- Does test material reflect diversity through the use of names, cultural references, pictures, and roles?

Differential Familiarity / Elitism

- Does the material contain phrases, concepts, and beliefs that are irrelevant to testing domain and are likely to be more familiar to specific groups than others?
- Does the material require knowledge of individuals, events, or groups that is not familiar to all groups of students?
- Does the material suggest that affluence is related to merit or intelligence?
- Does the material suggest that poverty is related to increased negative behaviors in society?
- Does the material use language, content, or context that is offensive to people of a particular economic status?
- Does success with the material assume that the examinee has experience with a certain type of family structure?
- Does the material favor one socioeconomic group over another?
- Does the material assume values not shared by all test takers?

Linguistic Features / Language Accessibility/Graphics

- Is grammar and vocabulary used in the items clear, concise and appropriate for the intended grade level?
- Are passages at a difficulty level that is appropriate for the intended grade level?
- Do the illustrations and graphics embody all of the previously referenced LABS Guidelines?

Other questions to consider

- Does the material favor one age group over others except in a context where experience or maturation is relevant?
- Does the material use language, content, or context that is not accessible to one or more of the age groups tested?
- Does the material contain language or content that contradicts values held by a certain culture?
- Does the material favor one racial or ethnic group over others?
- Does the material degrade people based on physical appearance or any physical, cognitive, or emotional challenge?
- Does the material focus only on a **person's** disability rather than portraying the whole person?
- Does the material favor one religion and/or demean others?

Exhibit 1: An Overview of Interaction Types

Interaction Types

An Overview

Multiple Choice

- One selection
- Any number of options
- Can orient options vertically, horizontally, stacked
- Very accessible

Multiple Select

- More than one selection by the student
- Can set a minimum/maximum number of options a student is required/allowed to select
- Any number of options
- Can orient options vertically, horizontally, stacked
- Very accessible

Table Match

- Within a table, student selects cells to match column headers with row headers
- Can set a minimum/maximum number of cells a student is required/allowed to select, based on
 - The whole table,
 - Each column, and/or
 - Each row
- Very accessible

Table Input

- Student inputs characters into cells of a table
- No rich text capabilities, only simple text
- Scoring is only robust for numbers
- Can set a validation for specific cells, e.g. only accept numeric characters typed into a table cell
- Very accessible

Edit Task

- Student can type in text to replace given text
- Given text is “crossed out” and replaced with student input
- Given text and student input text can only be simple text
- Very accessible

Edit Task Choice

- Student can choose an option to replace given text
- Given text is “crossed out” and replaced with student choice
- Given text and student choice can only be simple text
- Very accessible

Edit Task Choice Inline

- Student chooses text from a dropdown
- Rich Text (art, equation objects) is available
- Very accessible

Hot Text Selectable

- Student selects piece(s) of text and that text is highlighted
- Can set a minimum/maximum number of pieces of text the student is required/allowed to select
- Very accessible

Hot Text Reorderable

- Student is presented with blocks of text in a specific order that he or she can then reorder.
- Rich Text (art, equation objects) is available
- Can present vertically or horizontally oriented
- Not accessible, location-based

Hot Text Drag From Palette

- Student can select from a palette of options and drag an option into a gap within static text or in a table
- Rich text is available
- Only one option can be dragged into each gap
- Can set a minimum/maximum number of times a student is required/allowed to use a specific palette object
- Can orient the palette above or below the static text
- Not accessible – location based

Hot Text Custom

- Same as hot text item type in open office banks
- Encompasses the other three hot text interaction types, but with different rendering and different restrictions/limitations
- Generally, selectable hot text custom is accessible, while draggable is not

Text Entry

- Gives student text box to type into
- Used for handscored items or Natural Language
- Can set
 - Maximum length of student response in characters
 - Initial size (in terms of rows) of response box
 - Any editor that may be provided to the student (spell check, bolding, etc.)

Grid

- Allows student to drag/drop, draw dots or lines, and/or use hot spots
- Can use these functionalities together within one interaction
- Not accessible
 - Drag/Drop
 - Student drags images from a palette (or from the answer space itself) to different sections of the answer space
 - No control on how many images a student can drag to the answer space

Grid, ctd.

- Drawing
 - Student uses tools to create points, line segments, rays, and lines
 - Example: Washington Science 20092
- Hot Spot
 - Student selects predesignated regions of the answer space to shade, outline, or produce a predesignated image at a predesignated location

Graphic Gap Match

- Student drags an image or text from a palette to a predesignated location on the answer space
- Can set a minimum/maximum on the number of times a student is required/able to use each palette image/text
- Can set a minimum/maximum on the number of images/texts a student is required/able to drag to a specific location on the answer space
- Can orient palette horizontally or vertically, and to the top/bottom/left/right of the answer space

Hot Spot

- Student selects predesignated regions on the answer space, and that location is outlined
- Can set a minimum/maximum on the total number of regions a student is required/able to select
- Very selected response – most often there are other, better interactions

Equation

- Student uses a predesignated keypad to create mathematical objects, i.e. numbers, expressions, equations
- Scoring can look for specific responses (e.g., 2) or also equivalencies (e.g., 2 and $1 + 1$)
- Keyboard is accessible, but blind students must enter responses as text

Simulation

- Allow student to designate inputs in various ways
- An animation can be run with features based on those inputs
- An output table can be given to the student based on those inputs

Appendix F

Content Advisory Committee Participant Details

Table 1: Content Advisory Committee Participants, ELA

State	Date		Location	Grade/Grade Band	Number of Teachers in Each Group	Teacher Demographic Summary	Number of ELA Items Approved by Teacher Committees	Number of ELA Items Unique to State
Arizona	December	2014	Phoenix			Gender: Male 24%, Female 76% Ethnicity: White 66%, Asian 7%, African American 7%, Hispanic 7%, Native American 3%, Other 10% Region: Urban 72%, Suburban, 3%, Rural 25%	1,006	138
				3	6			
				4	7			
				5	7			
				6	8			
				7	6			
				8	6			
				9	5			
				10	5			
11	5							
Arizona	July	2015	Phoenix			Gender: Male 17%, Female 83% Ethnicity: White 83%, Biracial 4%, African American 4%, Hispanic 9% Region: Rural 5%, Suburban 4%, Urban 91%	601	114
				3-5	6			
				6-8	5			
				9	5			
				10	5			
11	5							
Arizona	July	2016	Phoenix			Gender: Male 17%, Female 83% Ethnicity: Asian 11%, Hispanic 11%, White 72%, Other 6% Region: Rural 6%, Suburban 6%, Urban 88%	550	69
				3-5	4			
				6-8	3			
				9	3			
				10	4			
11	4							

State	Date		Location	Grade/Grade Band	Number of Teachers in Each Group	Teacher Demographic Summary	Number of ELA Items Approved by Teacher Committees	Number of ELA Items Unique to State
Arizona	July	2017	Phoenix			Gender: Male 5%, Female 95% Ethnicity: Hispanic 22%, White 78% Region: Rural 12%, Urban 88%	469	223
				3-5	8			
				6-8	5			
				9-11	5			
Arizona	July	2018	Phoenix			Gender: Male 6%, Female 94%, Ethnicity: Asian 17%, Hispanic 22%, White 61% Region: Rural 11%, Urban 89%	469	266
				3-5	5			
				6-8	3			
				9-11	4			
Florida	October	2014	Jacksonville			Ethnicity: Asian 1%, African American 20%, Caucasian 68%, Hispanic 9%, Other 2% Gender: Female 76%, Male 24% Region: Panhandle 29%, East Central 16%, Northeast 17%, South 19%, West Central 19%	959	219
				4-7 (writing)	6			
				8-11 (writing)	6			
				5 (reading)				
	6-7 (reading)	8						
	8-9 (reading)	8						
	10-11 (reading)	8						
	September	2014	Jacksonville	4-5 (writing)	8			
6-7(writing)				8				
8-9(writing)				8				
10-11(writing)				8				
Florida	October	2015	Jacksonville			Ethnicity: Asian 1%, African American 20%, Caucasian 69%, Hispanic 8%, Other 2% Gender: Female 75%, Male 25% Region: Panhandle 32%, East Central 19%, Northeast 15%, South 17%, West Central 17%	444	153
				4-5	18			
				6-7	10			
				8	10			
				9	18			
				10	18			

State	Date		Location	Grade/Grade Band	Number of Teachers in Each Group	Teacher Demographic Summary	Number of ELA Items Approved by Teacher Committees	Number of ELA Items Unique to State
Florida	September	2016	Jacksonville			Ethnicity: Asian 1%, African American 20%, Caucasian 69%, Hispanic 8%, Other 2% Gender: Female 75%, Male 25% Region: Panhandle 32%, East Central 19%, Northeast 15%, South 17%, West Central 17%	243	26
				3-4	18			
				5-7	8			
				8-10	8			
Florida	November	2017	Jacksonville			Ethnicity: Asian 0%, African American 24%, Caucasian 66%, Hispanic 8%, Other 2% Gender: Female 74%, Male 22% Region: Panhandle 31%, East Central 20%, Northeast 18%, South 13%, West Central 18%	347	229
				3-4	6			
				4-5	6			
				6	6			
				7	6			
				8-9	7			
				9-10	5			
Utah	November	2014	Logan			Ethnicity: Native American 8%, Asian 8%, African American 8%, Hispanic 3%, Hawaiian/Pacific Islander 3%, White 70% Gender: Male 10%, Female 90% Teaching Experience: Regular Education 40%, Bilingual Education 3%, Special Education 3%, Administration 4%, Other 50%	595	24
				3-4	3			
				5-6	3			
				7-8	3			
				9-10	3			
				11	3			
Utah	September	2015	Provo			Gender: Male 13%, Female 87% Teaching Experience: Regular Education 87%, Special Education 13%	276	63
				3-4	4			
				5-6	5			
				7	4			
				8-9	3			
				10-11	4			

State	Date		Location	Grade/Grade Band	Number of Teachers in Each Group	Teacher Demographic Summary	Number of ELA Items Approved by Teacher Committees	Number of ELA Items Unique to State
Utah	September	2016	Provo			Ethnicity: White 82%, Black 6%, Asian 6%, American Indian 6% Gender: Female 100% Teaching Experience: Regular 50%, Bilingual Education 12%, Admin 19%, Other 37%	248	26
				3-6	6			
				7-8	3			
				9-11	6			
						Ethnicity: White 100% Gender: Female 100% Teaching Experience: Regular 63%, Other 25%. Blank 12%,		
				3-4	4			
				5-6	4			
				7	2			
		8	2					
Utah	August	2017	Salt Lake City			Ethnicity: White 100% Gender: Female 90%, Male 10% Teaching Experience: Regular 82%, Special Education 18%, Administration 18%, Other 18%	220	15
				3-5	6			
	6-8			6				
	September			3-5	6	Ethnicity: White 100% Gender: Female 100% Teaching Experience: Regular 75%, Administration 25%, Special Education 25%		
				6-8	6			
New Hampshire	November	2018	Meredith			Gender: Male 13%, Female 87% Teaching Experience: Regular Education 87%, Special Education 13%	159	159
				3-4	6			
				5-6	4			
	7-8	4						

State	Date		Location	Grade/Grade Band	Number of Teachers in Each Group	Teacher Demographic Summary	Number of ELA Items Approved by Teacher Committees	Number of ELA Items Unique to State
North Dakota	October	2018	Bismarck			Gender: Male 0%, Female 100% Region of the State: Northeast 21%, Northwest 5%, Southeast 47%, Southwest 26% Teaching Experience: English Language Learner 5%, Regular Education 79%, Special Education 16%	155	155
				3-5	5			
				6-8	5			
			HS	5				
West Virginia	November	2018	Charleston			Gender: Male 12%, Female 88% Ethnicity: White 88%, Asian 8%, African American 4% Teaching Experience: Regular Education 75%, Special Education 21%, Bilingual Education 4%,	111	111
				3-5	6			
			6-8	6				
Wyoming	October	2018	Cheyenne			Gender: Male 11%, Female 89% Teaching Experience: Regular Education Only 22%, Special Education 67%, English Language Learners 33% Gifted and Talented 33%	233	233
				3-5	6			
				6-8	6			
			HS	6				

Table 2: Content Advisory Committee Participants, Mathematics

State	Date		Location	Grade/Grade Band	Number of Teachers in Each Group	Teacher Demographic Summary	Number of Math Items Approved by Teacher Committees	Number of Math Items Unique to State
Arizona	December	2014	Phoenix			Gender: Male 7%, Female 93% Ethnicity: Asian 13%, African American 7%, White 73%, Other 7% Region: Rural 27%, Urban 73%	1,844	180
				Alg I	8			
				Alg II	7			
				Geometry	7			
				3	10			
				4	5			
				5	5			
				6	5			
				7	5			
8	5							
Arizona	July	2015	Phoenix			Gender: Male 23%, Female 77% Ethnicity: Asian 5%, Hispanic 5%, White 90% Region: Rural 4%, Urban 96%	270	128
				3-5	6			
				6-8	5			
				Alg I	5			
				Geometry	5			
Alg II	5							
Arizona	July	2016	Phoenix			Gender: Male 29%, Female 71% Ethnicity: White 57%, Asian 14%, Hispanic 29% Region: Suburban 14%, Urban 86%	522	144
				3-5	4			
				6-8	5			
				Alg I	4			
				Geometry	4			
Alg II	5							

State	Date		Location	Grade/Grade Band	Number of Teachers in Each Group	Teacher Demographic Summary	Number of Math Items Approved by Teacher Committees	Number of Math Items Unique to State
Arizona	July	2017	Phoenix			Gender: Male 19%, Female 89% Ethnicity: Hispanic 12%, White 84%, Multi-racial 4% Region: Rural 4%, Urban 96%	449	82
				3-5	8			
				6-8	9			
				Alg I	3			
				Geometry	3			
				Alg II	2			
Arizona	July	2018	Phoenix			Gender: Male 16%, Female 84% Ethnicity: Asian 11%, Hispanic 11%, White 78% Region: Rural 16%, Urban 84%	442	411
				Alg I	4			
				Alg II	3			
				Geometry	3			
				3-5	4			
				6-8	5			
Florida	October	2014	Jacksonville			Ethnicity: Asian 1%, African American 20%, Caucasian 68%, Hispanic 9%, Other 2% Gender: Female 76%, Male 24% Region: Panhandle 29%, East Central 16%, Northeast 17%, South 19%, West Central 19%	806	102
				5	8			
				6	8			
				7	8			
				8	8			
				Alg I	8			
				Geometry and Alg II	8			
Geometry	8							

State	Date		Location	Grade/Grade Band	Number of Teachers in Each Group	Teacher Demographic Summary	Number of Math Items Approved by Teacher Committees	Number of Math Items Unique to State
Florida	October	2015	Jacksonville			Ethnicity: Asian 1%, African American 20%, Caucasian 69%, Hispanic 8%, Other 2% Gender: Female 75%, Male 25% Region: Panhandle 32%, East Central 19%, Northeast 15%, South 17%, West Central 17%	519	90
				3	8			
				5 and 6	8			
				Geometry, Algebra II	8			
				4	8			
				7 and 8	8			
Geometry, Algebra 1	8							
Florida	September	2016	Jacksonville			Ethnicity: Asian 1%, African American 20%, Caucasian 69%, Hispanic 8%, Other 2% Gender: Female 75%, Male 25% Region: Panhandle 32%, East Central 19%, Northeast 15%, South 17%, West Central 17%	281	117
				3 and 5	5			
				6, 7, & 8	6			
				4 and 5	4			
Algebra 1, Algebra 2, Geometry	7							
Florida	November	2017	Jacksonville			Ethnicity: Asian 0%, African American 24%, Caucasian 66%, Hispanic 8%, Other 2% Gender: Female 74%, Male 22% Region: Panhandle 31%, East Central 20%, Northeast 18%, South 13%, West Central 18%	181	17
				Algebra 1	2 groups of 7			
				6 & 7	5			
				3 & 5	5			
				6 & 8	5			
				4 & 5	5			
Geometry	8							

State	Date		Location	Grade/Grade Band	Number of Teachers in Each Group	Teacher Demographic Summary	Number of Math Items Approved by Teacher Committees	Number of Math Items Unique to State
Utah	October	2014	Salt Lake City			Ethnicity: White 89%, Hispanic 4%, Native American 2%, African American 4% Gender: Female 74%, Male 26% Teaching Experience: Regular Education 68%, Bilingual 2%, Special Education 4%, no report 26%	544	24
				3-5	13			
				6-8	28			
Utah	August	2015	Provo			Ethnicity: White 93%, Hawaiian 3%, Native Hawaiian or Pacific Islander 3% Gender: Female 70%, Male 30% Teaching Experience: Regular Education 86%, Bilingual 3%, Special Education 6%, Administration 3%, Other 10%	603	63
				3-4	5			
				5	4			
				6	4			
				7	4			
				8	3			
				SM I	4			
				SM II	5			
				SM III	4			
Utah	August	2016	Park City			Ethnicity: White 94%, Blank 5% Gender: Female 76%, Male 23% Teaching Experience: Regular Education 94%, Bilingual 0%, Special Education 0%, Administration 0%, Other 5%	104	26
				3-6	6			
				3-5	6			
				6-8	6			
				7-11	6			

State	Date		Location	Grade/Grade Band	Number of Teachers in Each Group	Teacher Demographic Summary	Number of Math Items Approved by Teacher Committees	Number of Math Items Unique to State
Utah	July	2017	Salt Lake City			Ethnicity: White 100% Gender: Female 100% Teaching Experience: Regular Education 83%, Administration 17%	286	15
				3-5	6			
				6-8	6			
New Hampshire	November	2018	Meredith			Gender: Male 7%, Female 93% Teaching Experience: Regular Education 87%, Special Education 13%	98	84
				3-4	5			
				5-6	6			
				7-8	4			
North Dakota	October	2018	Bismarck			Gender: Male 11%, Female 89% Region of the State: Northeast 16%, Northwest 26%, Southeast 42%, Southwest 16% Teaching Experience: Regular Education 89%, Special Education 11%	164	141
				3-5	5			
				6-8	5			
				HS	5			
West Virginia	November	2018	Charleston			Gender: Male 12%, Female 88% Ethnicity: White 88%, Asian 8%, African American 4% Teaching Experience: Regular Education 75%, Special Education 21%, Bilingual Education 4%,	206	189
				3-5	6			
				6-8	6			
Wyoming	October	2018	Cheyenne			Gender: Male 17%, Female 83% Teaching Experience: Regular Education Only 44%, Regular and Special Education 56%	270	270
				3-5	6			
				6-8	5			
				HS	7			

Table 3: Content Advisory Committee and Fairness Committee Participants 2019-2020, ELA and Mathematics

State	Date		Location	Subject	Grade/Grade Band	Number of Teachers in Each Group	Teacher Demographic Summary	Number of Field Test Items Reviewed by Committees
West Virginia	July	2019	Charleston	ELA	3-5	6	Gender: Male 13%, Female 87% Ethnicity: White 92%, Asian 4%, African American 4%, Hispanic 0%, Native American 0%, Other 0% Region: Urban 31%, Suburban 4%, Rural 65%	307
				ELA	6-8	6		
				Math	3-5	6		
				Math	6-8	6		
				Fairness		6		
North Dakota	September	2019	Bismarck	ELA	3-5	5	Gender: Male 8%, Female 92% Ethnicity: White 91%, Asian 0%, African American 3%, Hispanic 0%, Native American 3%, Other 3% Region: Urban 27%, Suburban 8%, Rural 65%	365
				ELA	6-8	5		
				ELA	HS	5		
				Math	3-5	5		
				Math	6-8	5		
				Math	HS	5		
				Fairness		4		
				Fairness		5		
New Hampshire	October	2019	Meredith	ELA	3-5	4	Gender: Male 5%, Female 95% Ethnicity: White 100%, Asian 0%, African American 0%, Hispanic 0%, Native American 0%, Other 0% Region: Urban 15%, Suburban 25%, Rural 60%	32
				ELA	6-8	4		
				Math	3-5	4		
				Math	6-8	4		
				Fairness		5		
West Virginia	July	2020	Virtual	ELA	3-5	4	Gender: Male 9%, Female 91% Ethnicity: White 97%, Asian 0%, African American 3%, Hispanic 0%, Native American 0%, Other 0% Region: Urban 34%, Suburban 3%, Rural 63%	128
				ELA	6-8	4		
				Math	3-5	4		
				Math	6-8	4		
				Fairness		6		

Appendix G

Fairness Committee Participant Details

Table 1: Fairness Committee Participants

State	Subject Area	Date		Location	Teacher Demographic Summary by Year	Number of Items Reviewed		
Utah	ELA	September	2015	Provo	Ethnicity: White 18%, Native American 29%, Asian 12%, African American 15%, Hispanic 24%, Pacific Islander 2% Gender: Female 74%, Male 26% Teaching Experience: Regular Education 12%, Bilingual Education 0%, Special Education 12%, Administration 32%, Other 44%	3,796		
	Math	October		Provo				
Utah	ELA	November	2016	Provo				
	ELA	October		Provo				
	Math	November		Salt Lake City				
	Math	December		Provo				
Utah	ELA	September	2017	Salt Lake City			Ethnicity: White 18%, Native American 36%, Asian 9%, African American 9%, Hispanic 27%, Pacific Islander 0% Gender: Female 82%, Male 18% Teaching Experience: Regular Education 9%, Bilingual Education 0%, Special Education 9%, Administration 36%, Other 45%	575
	Math	August		Salt Lake City				
Florida	ELA and Math	September	2015	Jacksonville	Ethnicity: Asian 1%, African American 20%, Caucasian 69%, Hispanic 8%, Other 2% Gender: Female 75%, Male 25% Region: Panhandle 32%, East Central 19%, Northeast 15%, South 17%, West Central 17%	2,604		
Florida	ELA and Math	September	2016	Jacksonville				

State	Subject Area	Date		Location	Teacher Demographic Summary by Year	Number of Items Reviewed
Florida	ELA and Math	October	2017	Jacksonville	Ethnicity: Asian 1%, African American 20%, Caucasian 69%, Hispanic 8%, Other 2% Gender: Female 75%, Male 25% Region: Panhandle 32%, East Central 19%, Northeast 15%, South 17%, West Central 17%	3,724
Arizona	ELA and Math	September	2015	Phoenix	Gender: Male 20%, Female 80% Ethnicity: White 70%, Asian 7%, Hispanic 19%, African American 2%, Bi-racial 2% Region: Suburban 9%, Urban 89%, Rural 2%	
Arizona	ELA and Math	September	2016	Phoenix	Gender: Male 23%, Female 77% Ethnicity: White 65%, Asian 13%, Hispanic 20%, Other 2% Region: Suburban 20%, Urban 87%, Rural 3%	
Arizona	ELA and Math	September	2017	Phoenix	Gender: Male 12%, Female 88% Ethnicity: Hispanic 17%, White 81%, Multiracial 2% Region: Rural 8%, Urban 92%	
Arizona	ELA and Math	August	2018	Phoenix	Gender: Male 11%, Female 89% Ethnicity: Asian 14%, Hispanic 16%, White 70% Region: Rural 14%, Urban 86%	
New Hampshire	ELA and Math	November	2018	Meredith	Gender: Male 30%, Female 70%, Ethnicity: White 30%, Asian 20%, African American 10%, No Response 40% Teaching Experience: Special Education 10%, Bilingual Education 10%, Regular Education 80%	261

State	Subject Area	Date		Location	Teacher Demographic Summary by Year	Number of Items Reviewed
North Dakota	ELA and Math	October	2018	Bismarck	<p>Gender: Female 100% Male 0%</p> <p>Teaching Experience: Regular Education 100%, Special Education 0%, Bilingual Education 0%</p> <p>Region of the State: Northeast 25%, Northwest 25%, Southeast 25%, Southwest 25%</p>	340
West Virginia	ELA and Math	November	2018	Charleston	<p>Gender: Male 30%, Female 70%,</p> <p>Ethnicity: White 30%, Asian 20%, African American 10%, No Response 40%</p> <p>Teaching Experience: Special Education 10%, Bilingual Education 10%, Regular Education 80%</p>	853
Wyoming	ELA and Math	October	2018	Cheyenne	<p>Gender: Male 30%, Female 70%,</p> <p>Ethnicity: White 30%, Asian 20%, African American 10%, No Response 40%</p> <p>Teaching Experience: Special Education 10%, Bilingual Education 10%, Regular Education 80%</p>	507

Table 2: Content Advisory Committee and Fairness Committee Participants 2019-2020, ELA and Mathematics

State	Date		Location	Subject	Grade/Grade Band	Number of Teachers in Each Group	Teacher Demographic Summary	Number of Field Test Items Reviewed by Committees
West Virginia	July	2019	Charleston	ELA	3-5	6	Gender: Male 13%, Female 87% Ethnicity: White 92%, Asian 4%, African American 4%, Hispanic 0%, Native American 0%, Other 0% Region: Urban 31%, Suburban 4%, Rural 65%	307
				ELA	6-8	6		
				Math	3-5	6		
				Math	6-8	6		
				Fairness		6		
North Dakota	September	2019	Bismarck	ELA	3-5	5	Gender: Male 8%, Female 92% Ethnicity: White 91%, Asian 0%, African American 3%, Hispanic 0%, Native American 3%, Other 3% Region: Urban 27%, Suburban 8%, Rural 65%	365
				ELA	6-8	5		
				ELA	HS	5		
				Math	3-5	5		
				Math	6-8	5		
				Math	HS	5		
				Fairness		4		
				Fairness		5		
New Hampshire	October	2019	Meredith	ELA	3-5	4	Gender: Male 5%, Female 95% Ethnicity: White 100%, Asian 0%, African American 0%, Hispanic 0%, Native American 0%, Other 0% Region: Urban 15%, Suburban 25%, Rural 60%	32
				ELA	6-8	4		
				Math	3-5	4		
				Math	6-8	4		
				Fairness		5		
West Virginia	July	2020	Virtual	ELA	3-5	4	Gender: Male 9%, Female 91% Ethnicity: White 97%, Asian 0%, African American 3%, Hispanic 0%, Native American 0%, Other 0% Region: Urban 34%, Suburban 3%, Rural 63%	128
				ELA	6-8	4		
				Math	3-5	4		
				Math	6-8	4		
				Fairness		6		

Appendix H

Sample Data Review Training Materials

OHIO ASSESSMENT SYSTEM

Item Data Review

Item development process

- Initial item development cycle
- Fairness and sensitivity committee review
- Content advisory committee review
- Field test administration
- Final fairness and sensitivity review
- Item data review
- Form building

Field test goals

- Identify items that do not perform as intended
- Calibrate items to the bank scale

Field Testing

- Independent Field Tests
 - ▣ Ideal for field testing large number of items
- Embedded Field Tests
 - ▣ Ideal for bank replenishment
 - ▣ Operational test conditions

Field Test Analysis

- Classical statistics
 - Biserial correlations
 - p -values
 - Percent in response categories
- Differential Item Functioning
- Item Response Theory (IRT)

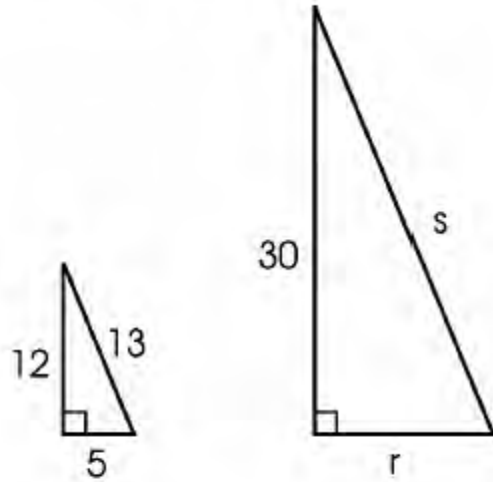
Statistical Review of Items

- Item Quality
 - ▣ Does the item behave the way it's supposed to behave?
- Item Difficulty
 - ▣ How hard is the item?
- Differential Item Functioning
 - ▣ Does the item behave differently across gender or major ethnic groups?

Item Cards

- Item Attributes
- Item Statistics
- Item Content

Two similar triangles are shown.



What are the correct lengths of sides r and s ?

- A. $r = 15, s = 39$
- B. $r = 23, s = 31$
- C. $r = 2.5, s = 5.2$
- D. $r = 12.5, s = 32.5$

Subject	Math
ITS ID	23517
Ohio Code	7M0000GXFXM3435D
Grade	7
Standard	GS
Benchmark	F
Indicator	5
Mathematical Process	N/A
Item Format	MC
Answer Key	D
Media Type	N/A

Item Statistics (FormD::G7M::SP13 -- Analysis Data: 908)		
Option	Percent	Correlation with Test
A	14.22%	-0.51
B	24.29%	-0.28
C	4.26%	-0.47
D	56.95%	0.63
Percent Omit	0.28%	
Fairness Statistics		
Female / Male		-A
Black / White		+A
Hispanic / White		-A
Multi-Race / White		+A

Item Quality

- Do highly skilled students perform better on the item than less skilled students?
- Correlation with Test – link between selecting a response option and doing well on the rest of the test
 - ▣ For key, + is good, - is bad
 - ▣ For distractors, - is good, + is bad

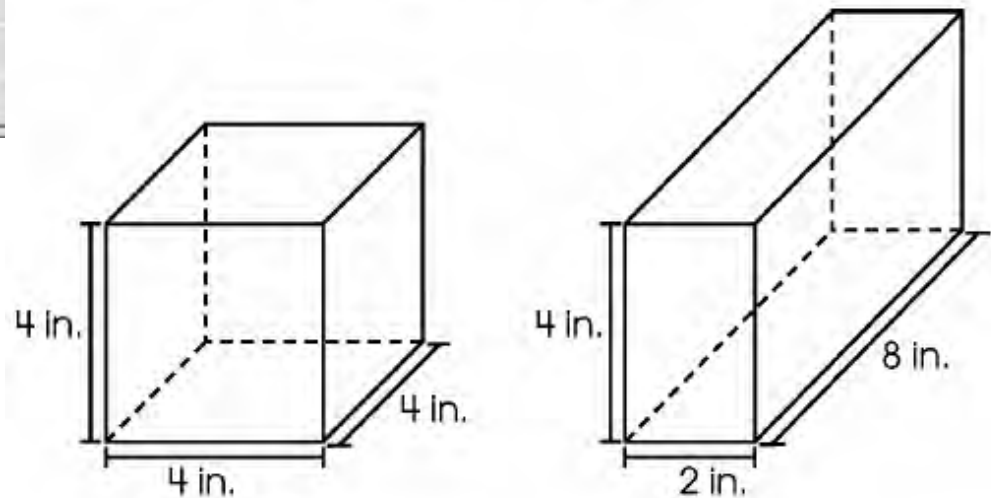
Subject	Math
ITS ID	18765
Ohio Code	7M0000MXGXM3090B
Grade	7
Standard	ME
Benchmark	G
Indicator	8
Mathematical Process	N/A
Item Format	MC
Answer Key	B
Media Type	Other

Item Statistics (FormA::G7M::SP13 -- Analysis Data: 905)

Option	Percent	Correlation with Test
A	10.21%	-0.18
B	63.26%	0.51
C	10.75%	-0.36
D	15.70%	-0.39
Percent Omit	0.08%	

Fairness Statistics

Two rectangular prisms and their dimensions are shown.



Which statement about these figures is true?

- A. They have the same volume and the same surface area.
- B. They have the same volume but different surface areas.
- C. They have different volumes but the same surface area.
- D. They have different volumes and different surface areas.

Subject	Math
ITS ID	23614
Ohio Code	7M0000AAKXM3455S
Grade	7
Standard	PA
Benchmark	K
Indicator	5
Mathematical Process	A
Item Format	SA
Answer Key	S
Media Type	N/A

Item Statistics (FormD::G7M::SP13 -- Analysis Data: 908)		
Points	Percent in Category	Average Score of Students in Category
2	28.07%	39.20
1	24.20%	31.51
0	43.67%	23.13
Percent Omit	4.07%	
Correlation with Test		0.66
Proportion Correct	40.17%	
Fairness Statistics		
Female / Male		+B
Black / White		+A
Hispanic / White		+A
Multi-Race / White		+A

In your Answer Document, create a table with three ordered pairs that can be used to graph the equation $y = 2x + 3$.

Then, draw a line on a coordinate grid that represents this equation.

Subject	Science	Item Statistics (FormG::G8S::SP13 -- Analysis Data: 928)		
ITS ID	23891	Option	Percent	Correlation with Test
Ohio Code	8S0000ESDXC2118D	A	40.52%	0.35*
Test Level	8	B	13.40%	-0.16
Grade	6	C	19.62%	-0.24
Standard	ES	D	26.27% *	-0.10*
Content Standards	ES	Percent Omit	0.19%	
Benchmark	D	Fairness Statistics		
Indicator	1	Female / Male		+A
Item Format	MC	Black / White		+B
Answer Key	D	Hispanic / White		+B
Media Type	N/A	Multi-Race / White		+B

A rock is composed of coarse-grained, light and dark minerals that are in layers or bands.

What can be determined about the environment in which this rock formed?

- The rock formed in a riverbed.
- The rock formed at a mid-Atlantic ridge.
- The rock formed at the surface of a volcano.
- The rock formed underground near a volcanic mountain range.

Subject	Math	Item Statistics (FormD::G7M::SP13 -- Analysis Data: 908)		
ITS ID	21692	Option	Percent	Correlation with Test
Ohio Code	7M0000DXEXM3307C	A	2.03%	-0.44
Grade	7	B	4.80%	-0.46
Standard	DA	C	33.67% *	-0.05*
Benchmark	E	D	59.35%	0.23*
Indicator	5	Percent Omit	0.15%	
Mathematical Process	N/A	Fairness Statistics		
Item Format	MC	Female / Male		+B
Answer Key	C	Black / White		-A
Media Type	N/A	Hispanic / White		-A
		Multi-Race / White		-A

Juan wants to determine the most popular movie among seventh-graders at his school.

Which survey method will result in a sample that most accurately reflects the opinions of all seventh-graders at the school?

- Ask all seventh-graders on his school bus.
- Survey the first 50 students from the alphabetical list of all seventh-graders who attend the school.
- Select 50 students at random from all seventh-graders who attend the school.
- Distribute a survey form to all seventh-graders in his school and ask them to return the completed form on the following day.

Subject	Math	Item Statistics (FormA::G4M::SP13 -- Analysis Data: 1360)		
ITS ID	19973	Option	Percent	Correlation with Test
Ohio Code	4M0000DXAXL1118C	A	17.14%	-0.24
Grade	4	B	21.70%	-0.05
Standard	DA	C	51.01%	0.20*
Benchmark	A	D	9.83%	-0.01
Indicator	1	Percent Omit	0.33%	
Mathematical Processes	N/A	Fairness Statistics		
Item Format	MC	Female / Male		+B
Answer Key	C	Black / White		-A
		Hispanic / White		-A
		Multi-Race / White		+A

Maria wants to find out which animal the students in her school want as a new mascot.

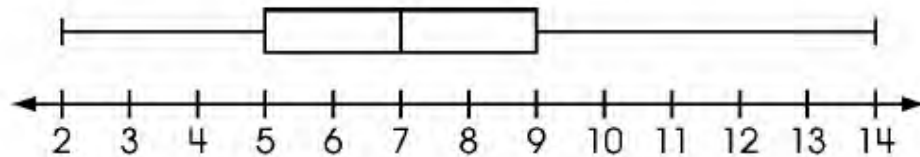
Which group of students should Maria survey to answer her question?

- every fourth-grade girl in the school
- every student riding Maria's school bus
- every tenth student from a list of all students
- every fifth student in line to buy lunch in the cafeteria

Subject	Math
ITS ID	21699
Ohio Code	7M0000DXAXM3314C
Grade	7
Standard	DA
Benchmark	A
Indicator	1
Mathematical Process	N/A
Item Format	MC
Answer Key	C
Media Type	Picture/Graphic

Item Statistics (FormB::G7M::SP13 -- Analysis Data: 906)		
Option	Percent	Correlation with Test
A	7.47%	-0.32
B	42.88%	0.04
C	28.62% *	0.10*
D	20.90%	-0.01
Percent Omit	0.13%	
Fairness Statistics		

Keisha surveyed her classmates about the number of books they read over the summer. She displayed the results in the box-and-whisker plot shown.



Number of Books Read

Keisha claims that the number of students who read 5 or fewer books is equal to the number of students who read 9 or more books.

Which statement best supports Keisha's claim?

- A. The range of the data is 2 to 14.
- B. The median is exactly halfway between 5 and 9.
- C. The first quartile is 2 to 5 and the fourth quartile is 9 to 14.
- D. The data has half of the values less than 7 and half of the values greater than 7.

Item Difficulty

- How hard is the item?
- What percent of students answer item correctly?

Subject	Math
ITS ID	7754
Ohio Code	6M0000NXIXM1126D
Grade	6
Standard	NS
Benchmark	I
Indicator	14
Mathematical	N/A

Item Statistics (FormA::G6M::SP13 -- Analysis Data: 891)		
Option	Percent	Correlation with Test
A	10.80%	-0.40
B	4.00%	-0.21
C	53.66%	-0.18
D	31.47% *	0.46
Percent Omit	0.07%	

Process	Subject	Math
Item Form	ITS ID	20324
Answer	Ohio Code	5M0000AXDXM1514B
Media T	Grade	5
	Standard	PA
	Benchmark	D
	Indicator	3
Mathematical	N/A	

Item Statistics (FormA::G5M::SP13 -- Analysis Data: 1371)		
Option	Percent	Correlation with Test
A	40.82%	-0.37
B	48.69%	0.58
C	6.85%	-0.44
D	3.46%	-0.37
Percent Omit	0.18%	

Process	Subject	Science
Item Form	ITS ID	23697
Answer Key	Ohio Code	5S0000ESBXR3182C
	Test Level	5
	Grade	4
	Standard	ES
	Content Standard	ES
	Benchmark	B
	Indicator	8
	Item Format	MC
	Answer Key	C
	Media Type	N/A

Item Statistics (FormA::G5S::SP13 -- Analysis Data: 1344)		
Option	Percent	Correlation with Test
A	8.17%	-0.46
B	6.62%	-0.53
C	77.08%	0.59
D	7.99%	-0.26
Percent Omit	0.14%	
Fairness Statistics		
Female / Male		+A
Black / White		-B
Hispanic / White		-B
Multi-Race / White		-B

Subject	Math
ITS ID	10780
Ohio Code	5M0000AXAXM0500S
Grade	5
Standard	PA
Benchmark	A
Indicator	1
Mathematical Processes	N/A
Item Format	SA
Answer Key	S

Item Statistics (FORMB::G5M::SP13 -- Analysis Data: 1341)		
Points	Percent in Category	Average Score of Students in Category
2	53.33%	43.47
1	36.13%	33.77
0	10.40%	22.83
Percent Omit	0.13%	
Correlation with Test		0.61
Proportion Correct	71.40%	
Fairness Statistics		

Subject	Science
ITS ID	24103
Ohio Code	5S0000ESCXC3288S
Test Level	5
Grade	5
Standard	ES
Content Standard	ES
Benchmark	C
Indicator	5
Item Format	SA
Answer Key	S
Media Type	N/A

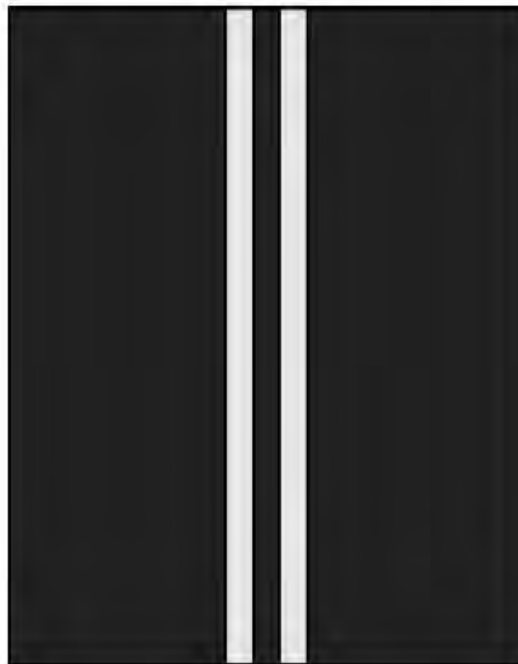
Item Statistics (FormB::G5S::SP13 -- Analysis Data: 1345)		
Points	Percent in Category	Average Score of Students in Category
2	14.60%	41.90
1	34.07%	37.68
0	51.00%	29.45
Percent Omit	0.33%	
Correlation with Test		0.52
Proportion Correct	31.63%	
Fairness Statistics		
Female / Male		-A
Black / White		-B
Hispanic / White		-A
Multi-Race / White		-A

Subject	Math
ITS ID	18206
Ohio Code	5M0000GXDXL1486B
Grade	5
Standard	GS
Benchmark	D
Indicator	2
Mathematical Processes	N/A
Item Format	MC
Answer Key	B

Item Statistics (FORMC::G5M::SP13 -- Analysis Data: 1342)

Option	Percent	Correlation with Test
A	0.95%	-0.44
B	96.09%*	0.46
C	2.14%	-0.39
D	0.62%	-0.38
Percent Omit	0.20%	

The painted lines on the road shown are exactly five inches apart.



Which word describes the relationship of the painted lines on the road?

- A. intersecting
- B. parallel
- C. perpendicular
- D. skew

Subject	Math
ITS ID	22812
Ohio Code	6M0000NXIXM3349B
Grade	6
Standard	NS
Benchmark	I
Indicator	15
Mathematical Process	N/A
Item Format	MC
Answer Key	B
Media Type	N/A

Item Statistics (FormB::G6M::SP13 -- Analysis Data: 892)		
Option	Percent	Correlation with Test
A	60.54%	-0.18
B	16.66%**	0.34
C	11.53%	-0.14
D	11.22%	0.06*
Percent Omit	0.05%	
Fairness Statistics		
Female / Male		-B
Black / White		+A
Hispanic / White		+A
Multi-Race / White		+A

Selima bought a dress for \$80 from a warehouse and sold it for \$100 at her clothing store.

What percent markup did Selima use?

- A. 20%
- B. 25%
- C. 80%
- D. 125%

Differential Item Functioning (DIF)

- Fair Items behave similarly across groups
- Probability of answering correctly is the same for all students of similar ability regardless of group membership
- Group comparisons
 - ▣ Black/African American vs. White
 - ▣ Hispanic vs. White
 - ▣ Multi-ethnic vs. White
 - ▣ Female vs. Male

No Differential Item Functioning

	Item Difficulty (p -value) by Ability Level				
Ethnicity	Bottom Quintile	2 nd Quintile	3 rd Quintile	4 th Quintile	Top Quintile
Black/African American	.20	.35	.50	.65	.80
White	.20	.35	.50	.65	.80

Differential Item Functioning

	Item Difficulty (p -value) by Ability Level				
Ethnicity	Bottom Quintile	2 nd Quintile	3 rd Quintile	4 th Quintile	Top Quintile
Black/African American	.05	.20	.35	.50	.65
White	.20	.35	.50	.65	.80

DIF Classifications

- Direction of possible bias
 - “-” item favors whites/males
 - “+” item favors focal group
- Severity of possible bias
 - “A” No statistical evidence of DIF
 - “B” Evidence for potential mild DIF
 - “C” Evidence for potential severe DIF

A biologist counts the number of flowers growing in a greenhouse each year. The data for the first 4 years are shown in the table.

Year	Number of Flowers
1	20
2	60
3	180
4	540
5	?

If the pattern continues, how many flowers will be growing in the greenhouse in year 5?

- A. 580
- B. 940
- C. 1080
- D. 1620

Benchmark	A
Indicator	1
Mathematical Process	N/A
Item Format	MC
Answer Key	D
Media Type	Table

::SP13 -- Analysis Data: 906)		
	Percent	Correlation with Test
C	11.01%	-0.30
D	70.89%	0.59
Percent Omit	0.10%	
Fairness Statistics		
Female / Male		-C*
Black / White		-B
Hispanic / White		-A
Multi-Race / White		-A

Subject	Science	Item Statistics (FormH::G8S::SP13 -- Analysis Data: 929)		
ITS ID	23909	Option	Percent	Correlation with Test
Ohio Code	8S0000PSCXC2131A	A	26.00% *	0.45
Test Level	8	B	38.85%	-0.22
Grade	6	C	2.29%	-0.43
Standard	PS	D	32.80%	-0.11
Content Standards	PS	Percent Omit	0.06%	
Benchmark	C	Fairness Statistics		
Indicator	6	Female / Male		-B
Item Format	MC	Black / White		+B
Answer Key	A	Hispanic / White		+A
Media Type	N/A	Multi-Race / White		+A

Subject	Science	Item Statistics (FormD::G8S::SP13 -- Analysis Data: 925)		
ITS ID	23909	Option	Percent	Correlation with Test
Ohio Code	8S0000PSCXC2131A	A	25.81% *	0.47
Test Level	8	B	38.25%	-0.22
Grade	6	C	2.42%	-0.49
Standard	PS	D	33.48%	-0.11
Content Standards	PS	Percent Omit	0.03%	
Benchmark	C	Fairness Statistics		
Indicator	6	Female / Male		-C*
Item Format	MC	Black / White		+B
Answer Key	A	Hispanic / White		-A
Media Type	N/A	Multi-Race / White		+A

Why is wind energy considered a form of solar energy?

- A. Heat from the sun causes wind.
- B. Both are natural forms of energy.
- C. Wind is only present during the day.
- D. Both are forms of renewable energy.

Expert Judges

- Statistical information is important, but not a substitute for expert judges
- Items may show DIF because some concepts may be less likely to be covered in low income area schools

Classical Item Flags

- Correlation with Total Test Score
 - ▣ Flagged if less than .25
 - ▣ For distractor responses in MC items, flagged if greater than .05
- Percent Selecting Response Option (MC Items)
 - ▣ For keyed responses, flagged if less than 25% (too hard) or greater than 95% (too easy)
- Non-Modal Key Response
 - ▣ MC items flagged if keyed response is not modal student response
- Percent Omitted
 - ▣ Flagged if greater than 10%

DIF Flag

- Fairness Statistics
 - “C” indicates that the item is more difficult for one group and should be reviewed carefully for bias
- Direction of possible bias
 - “-” item favors whites/males
 - “+” item favors focal group
- Severity of possible bias
 - “A” No statistical evidence of DIF
 - “B” Evidence for potential mild DIF
 - “C” Evidence for potential severe DIF

Appendix I

Data Review Committee Participant Details

Table 1: Data Review Committee Participants

State	Date		Location	Grade/Grade Band	Teacher Demographic Summary by Year	Number of Items Reviewed
Arizona	July	2017	Phoenix		Gender: Male 7%, Female 93% Ethnicity: Hispanic 29%, White 71% Region: Rural 8%, Urban 92%	1,072
				ELA 3-6		
				ELA 7-11		
				Math 3-6		
				Math 7-8		
				Alg I, II and Geometry		
Arizona	July	2018	Phoenix		Gender: Male 9%, Female 91% Ethnicity: Asian 9%, Hispanic 18%, White 73% Region: Rural 19%, Urban 81%	918
				ELA 3-5		
				ELA 3-5		
				ELA 6-8		
				ELA 9-11		
				Math 3-5		
				Math 6-8		
				Alg I		
				Alg II		
				Geometry		
Utah	July	2015	Provo		Ethnicity: White 100% Gender: Female 93%, Male 7% Teaching Experience: Regular Education 50%, Bilingual Education 14%, Administration 21%, Other 28%	1,139
				Math 3-5		
				Math 6-8		
				Secondary Math		
				ELA 3-5		
				ELA 6-8		
				ELA 9-11		

State	Date		Location	Grade/Grade Band	Teacher Demographic Summary by Year	Number of Items Reviewed
Utah	July	2016	Provo		Ethnicity: White 96%, Asian 2%, Other 2%, Gender: Male 7%, Female 93% Teaching Experience: Regular Education 95%, Special Education 5%, Bilingual Education 0%	879
				Math 3-5		
				Math 6-8		
				Secondary Math		
				ELA 3-5		
				ELA 6-8		
ELA 9-11						
Utah	July	2017	Provo		Ethnicity: White 92%, Hispanic 3%, Other 5% Gender: Male 11%, Female 89% Teaching Experience: Regular Education 97%, Special Education 3%, Bilingual Education 0%	352
				Math 3-5		
				Math 6-8		
				Secondary Math		
				ELA 3-5		
				ELA 6-8		
ELA 9-11						

Appendix J

Test Form Review Committee Participant Details

Table 1: Test Form Review Committee Participants, ELA

State	Date		Location	Grade/Grade Band	Number of Teachers in Each Group	Teacher Demographic Summary
West Virginia	November	2018	Charleston			Gender: Male 12%, Female 88% Ethnicity: White 88%, Asian 8%, African American 4% Teaching Experience: Regular Education 75%, Special Education 21%, Bilingual Education 4%,
				3-5	6	
				6-8	6	

Table 2: Test Form Review Committee Participants, Mathematics

State	Date		Location	Grade/Grade Band	Number of Teachers in Each Group	Teacher Demographic Summary
West Virginia	November	2018	Charleston			Gender: Male 12%, Female 88% Ethnicity: White 88%, Asian 8%, African American 4% Teaching Experience: Regular Education 75%, Special Education 21%, Bilingual Education 4%,
				3-5	6	
				6-8	6	

Appendix K

ICCR Adaptive Algorithm Design

TABLE OF CONTENTS

1. INTRODUCTION, BACKGROUND, AND DEFINITIONS.....	2
1.1 Blueprint.....	3
1.2 Content Value.....	4
1.2.1 Content Value for Single Items.....	4
1.2.2 Content Value for Sets of Items.....	5
1.3 Information Value.....	6
1.3.1 Individual Information Value.....	7
1.3.2 Binary Items.....	7
1.3.3 Polytomous Items.....	7
1.3.4 Item Group Information Value.....	10
2. ENTRY AND INITIALIZATION.....	10
2.1 Item Pool.....	10
2.2 Adjust Segment Length.....	10
2.3 Initialization of Starting Theta Estimates.....	11
2.4 Insertion of Embedded Field-Test Items.....	11
3. ITEM SELECTION.....	12
3.1 Trimming the Custom Item Pool.....	13
3.2 Recycling Algorithm.....	14
3.3 Adaptive Item Selection.....	14
3.4 Selection of the Initial Item.....	15
3.5 Exposure Control.....	15
4. TERMINATION.....	15
A1. DEFINITIONS OF USER-SETTABLE PARAMETERS.....	16
A2. SUPPORTING DATA STRUCTURES.....	17

ICCR ADAPTIVE ITEM SELECTION ALGORITHM

1. INTRODUCTION, BACKGROUND, AND DEFINITIONS

This document describes the ICCR adaptive item selection algorithm. The item selection algorithm is designed to cover a standards-based blueprint, which may include content, cognitive complexity, and item type constraints. The item selection algorithm will also include:

- the ability to customize an item pool based on access constraints and screen items that have been previously viewed or may not be accessible for a given individual;
- a mechanism for inserting embedded field-test items; and
- a mechanism for delivering “segmented” tests in which separate parts of the test are administered in a fixed order.

This document describes the algorithm and the design for its implementation for the ICCR Test Delivery System. The implementation builds extensively on the algorithm implemented in CAI’s Test Delivery System and incorporates substantial CAI intellectual property. CAI will release the algorithm and the implementation described here under the same open-source license under which the rest of the open-source system is released.

The general approach described here is based on a highly parameterized multiple-objective utility function. The objective function includes:

- a measure of content match to the blueprint;
- a measure of overall test information; and
- measures of test information for each reporting category on the test.

We define an objective function that measures an item’s contribution to each of these objectives, weighting them to achieve the desired balance among them. Equation 1 sketches this objective function for a single item.

$$f_{ijt} = w_2 \sum_{r=1}^R s_{rit} p_r d_{rj} + w_1 \sum_{k=1}^K q_k h_k(v_{kijt}, V_{kit}, t_k) + w_0 h_0(u_{ijt}, U_{it}, t_0) \quad (1)$$

where the term w represents user-supplied weights that assign relative importance to meeting each of the objectives, d_{rj} indicates whether item j has the blueprint-specified feature r , and p_r is the user-supplied priority weight for feature r . The term s_{rit} is an adaptive control parameter that is described below. In general, s_{rit} increases for features that have not met their designated minimum as the end of the test approaches.

The remainder of the terms represents an item’s contribution to measurement precision:

- v_{kijt} is the value of item j toward reducing the measurement error for reporting category k for examinee i at selection t ; and

- u_{ijt} is the value of item j in terms of reducing the overall measurement error for examinee i at selection t .

The terms U_{it} and V_{kit} represent the total information overall and on reporting category k , respectively.

The term q_k is a user-supplied priority weight associated with the precision of the score estimate for reporting category k . The term t represents precision targets for the overall score (t_0) and each score reporting category score. The functions $h(\cdot)$ are given by:

$$h_0(u_{ijt}, U_{it}, t_0) = \begin{cases} au_{ijt} & \text{if } U_{it} < t_0 \\ bu_{ijt} & \text{otherwise} \end{cases}$$

$$h_{1k}(v_{kijt}, V_{kit}, t_k) = \begin{cases} c_k v_{kijt} & \text{if } V_{kit} < t_k \\ d_k v_{kijt} & \text{otherwise} \end{cases}$$

Items can be selected to maximize the value of this function. This objective function can be manipulated to produce a pure, standards-free adaptive algorithm by setting w_2 to zero or a completely blueprint-driven test by setting $w_1 = w_0 = 0$. Adjusting the weights to optimize performance for a given item pool will enable users to maximize information subject to the constraint that the blueprint is virtually always met.

We note that the computations of the content values and information values generate values on very different scales and that the scale of the content value varies as the test progresses. Therefore, we normalize both the information and content values before computing the value of Equation 1. This

normalization is given by $x = \begin{cases} 1 & \text{if } \min = \max \\ \frac{v - \min}{\max - \min} & \text{otherwise} \end{cases}$, where \min and \max represent the minimum and maximum, respectively, of the metric computed over the current set of items or item groups.

The remainder of this section describes the overall program flow, the form of the blueprint, and the various value calculations employed in the objective function. Subsequent sections describe the details of the selection algorithm.

1.1 Blueprint

Each test will be described by a single blueprint for each segment of the test and will identify the order in which the segments appear. The blueprint will include:

- an indicator of whether the test is adaptive or fixed form;
- termination conditions for the segment, which are described in a subsequent section;
- a set of nested content constraints, each of which is expressed as:
 - the minimum number of items to be administered within the content category;
 - the maximum number of items to be administered within the content category;
 - an indication of whether the maximum should be deterministically enforced (a “strict” maximum);

- a priority weight for the content category p_r ;
- an explicit indicator as to whether this content category is a reporting category; and
- an explicit precision-priority weight (q_k) for each group identified as a reporting category.
- a set of non-nested content constraints, which are represented as:
 - a name for the collection of items meeting the constraint;
 - the minimum number of items to be administered from this group of items;
 - the maximum number of items to be administered from this group of items;
 - an indication of whether the maximum should be deterministically enforced (a “strict” maximum);
 - a priority weight for the group of items p_r ;
 - an explicit indicator as to whether this named group will make up a reporting category; and
 - an explicit precision-priority weight (q_k) for each group identified as a reporting category.
- The priority weights, p_r on the blueprint, can be used to express values in the blueprint match. Large weights on reporting categories paired with low (or zero) weights on the content categories below them may allow more flexibility to maximize information in a content category covering fewer fine-grained targets, while the reverse would mitigate toward more reliable coverage of finer-grained categories, with less content flexibility within reporting categories.

An example of a blueprint specification appears in Appendix 1.

Each segment of a test will have a separate blueprint.

1.2 Content Value

Each item or item group will be characterized by its contribution to meeting the blueprint, given the items that have already been administered at any point. The contribution is based on the presence or absence of features specified in the blueprint and denoted by the term d in Equation 1. This section describes the computation of the content value.

1.2.1 Content Value for Single Items

For each constraint appearing in the blueprint (r), an item i either does or does not have the characteristic described by the constraint. For example, a constraint might require a minimum of four and a maximum of six algebra items. An item measuring algebra has the described characteristic, and an item measuring geometry but algebra does not. To capture this constraint, we define the following:

- d_i is a feature vector in which the elements are d_{ir} , summarizing item i 's contribution to meeting the blueprint. This feature vector includes content categories such as claims and targets as well as other features of the blueprint, such as Depth of Knowledge and item type.

- S_{it} is a diagonal matrix, the diagonal elements of which are the adaptive control parameters s_{rit} .
- p is the vector containing the user-supplied priority weights p_r .

The scalar content value for an item is given by $C_{ijt} = d_i' S_{it} p$.

Letting z_{rit} represent the number of items with feature r administered to student i by iteration t , the value of the adaptive control parameters is:

$$s_{rit} = \begin{cases} m_{it} \left(2 - \frac{z_{rit}}{Min_r} \right) & \text{if } z_r < Min_r \\ 1 - \frac{z_{rit} - Min_r}{Max_r - Min_r} & \text{if } Min_r < z_{rit} < Max_r \\ (Max_r - z_{rit}) - 1 & \text{if } Max_r \leq z_{rit} \end{cases}$$

The blueprint defines the minimum (Min_r) and maximum (Max_r) number of items to be administered with each characteristic (r).

The term $m_{it} = \frac{T}{T-t}$ where T is the total test length. This has the effect of increasing the algorithm's preference for items that have not yet met their minimums as the end of the test nears and the opportunities to meet the minimum diminish.

This increases the likelihood of selecting items for content that has not met its minimum as the opportunities to do so are used up. The value s is highest for items with content that has not met its minimum, declines for items representing content for which the minimum number of items has been reached but the maximum has not, and turns negative for items representing content that has met the maximum.

1.2.2 Content Value for Sets of Items

Calculation of the content value of sets of items is complicated by two factors:

1. The desire to allow more items to be developed for each set and to have the most advantageous set of items administered
2. The design objective of characterizing the information contribution of a set of items as the expected information over the working theta distribution for the examinee

The former objective is believed to enhance the ability to satisfy highly constrained blueprints while still adapting to obtain good measurement for a broad range of students. The latter arises from the recognition that ELA tests will select one set of items at a time, without an opportunity to adapt once the passage has been selected.

The general approach involves successive selection of the highest content value item in the set until the indicated number of items in the set have been selected. Because the content value of an item changes with each selection, a temporary copy of the already-administered content vector for the examinee is updated with each selection such that subsequent selections reflect the items selected in previous iterations.

Exhibit 1 presents a flowchart for this calculation. Readers will note the check to determine whether $w_0 > 0$ or $w_1 > 0$. These weights, defined with Equation 1, identify the user-supplied importance of information optimization relative to blueprint optimization. In cases such as independent field tests, this weight may be set to zero, as it may not be desirable to make item administration dependent on match to student performance. In more typical adaptive cases where item statistics will not be recalculated, favoring more informative items is generally better. The final measure of content value for the set of selected set of items is divided by the number of items selected to avoid a bias toward selection of sets with more items.

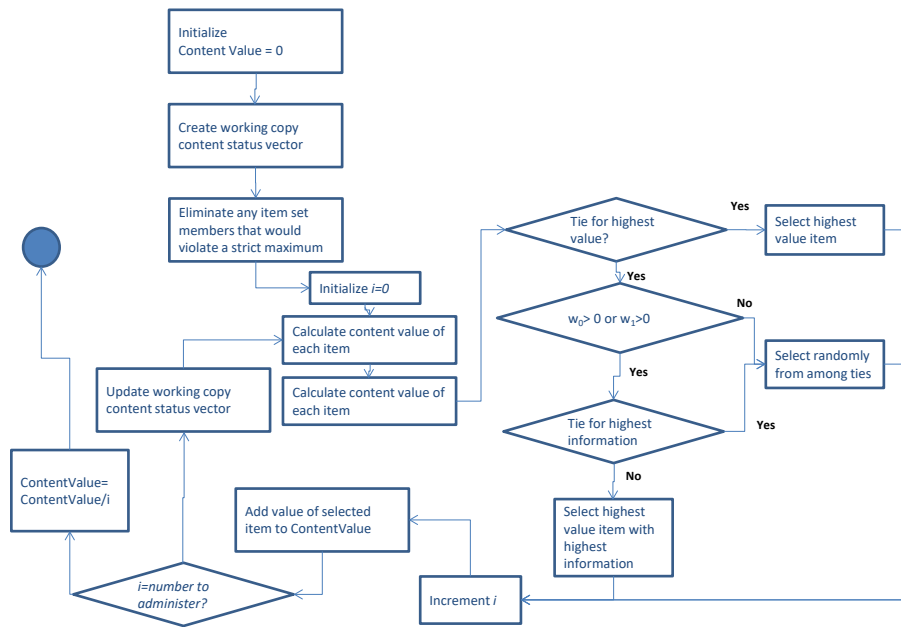


Exhibit 1. Content Value Calculation for Item Sets

1.3 Information Value

Each item or item group also has value in terms of maximizing information, both overall and on reporting categories.

1.3.1 Individual Information Value

The information value associated with an item will be an approximation of information. The system will be designed to use generalized IRT models; however, it will treat all items as though they offer equal measurement precision. This is the assumption made by the Rasch model, but in more general models, items known to offer better measurement are given preference by many algorithms. Subsequent algorithms are then required to control the exposure of the items that measure best. Ignoring the differences in slopes serves to eliminate this bias and help equalize exposure.

1.3.2 Binary Items

The approximate information value of a binary item will be characterized as $I_j(\theta) = p_j(\theta)(1 - p_j(\theta))$, where the slope parameters are artificially replaced with a constant.

1.3.3 Polytomous Items

In terms of information, the best polytomous item in the pool is the one that maximizes the expected information, $I_j(\theta)$. Formally, $I_j(\theta) > I_k(\theta)$ for all items $k \neq j$. The true value θ , however, remains unknown and is accessed only through an estimate, $\hat{\theta} \sim N(\bar{\theta}, \sigma_\theta)$. By definition of an expectation, the expected information $I_j(\theta) = \int I_j(t)f(t|\bar{\theta}, \sigma_\theta)dt$.

The intuition behind this result is illustrated in Exhibit 2. In Exhibit 2, each panel graphs the distribution of the estimate of θ for an examinee. The top panel assumes a polytomous item in which one step threshold (A1) matches the mean of the θ estimate distribution. In the bottom panel, neither step threshold matches the mean of the θ estimate distribution. The shaded area in each panel indicates the region in which the hypothetical item depicted in the panel provides more information. We see that approximately 2/3 of the probability density function is shaded in the lower panel, while the item depicted in the upper panel dominates in only about 1/3 of the cases. In this example, the item depicted in the lower panel has a much greater probability of maximizing the information from the item, despite the fact that the item in the upper panel has a threshold exactly matching the mean of the estimate distribution and the item in the lower panel does not.

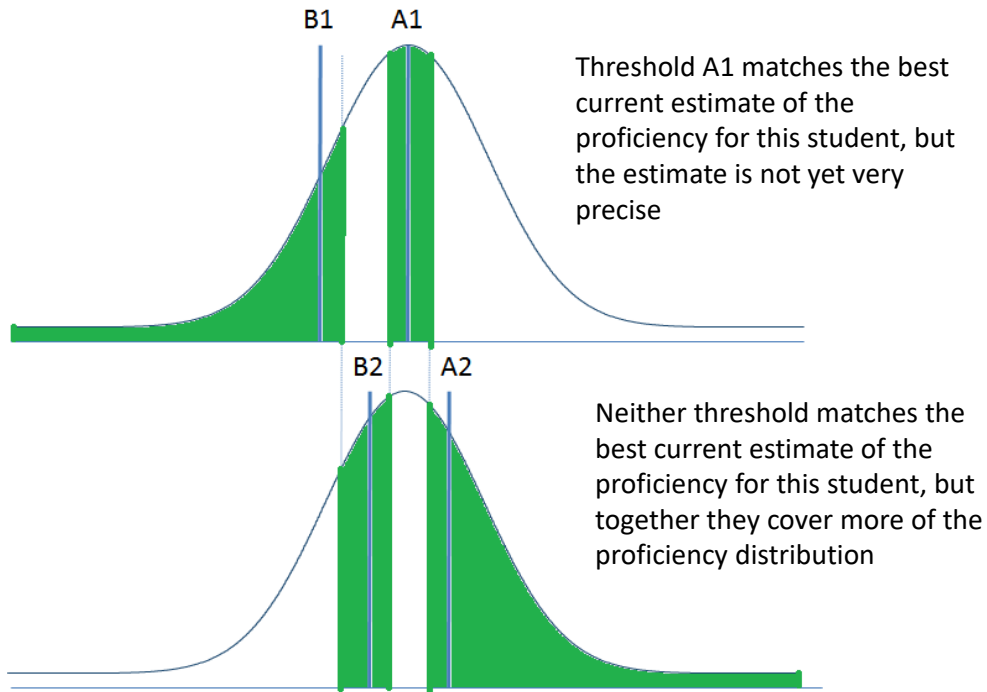


Exhibit 2. Two example items, with the shaded region showing the probability that the item maximizes information for the examinee depicted.

Exhibit 3 shows what happens to information as the estimate of this student's proficiency becomes more precise (later in the test). In this case, the item depicted in the top panel maximizes information about 65-70 percent of the time, compared to about 30 to 35 percent for the item depicted in the lower panel. These are the same items depicted in the Exhibit 2, but in this case we are considering information for a student with a more precise current proficiency estimate.

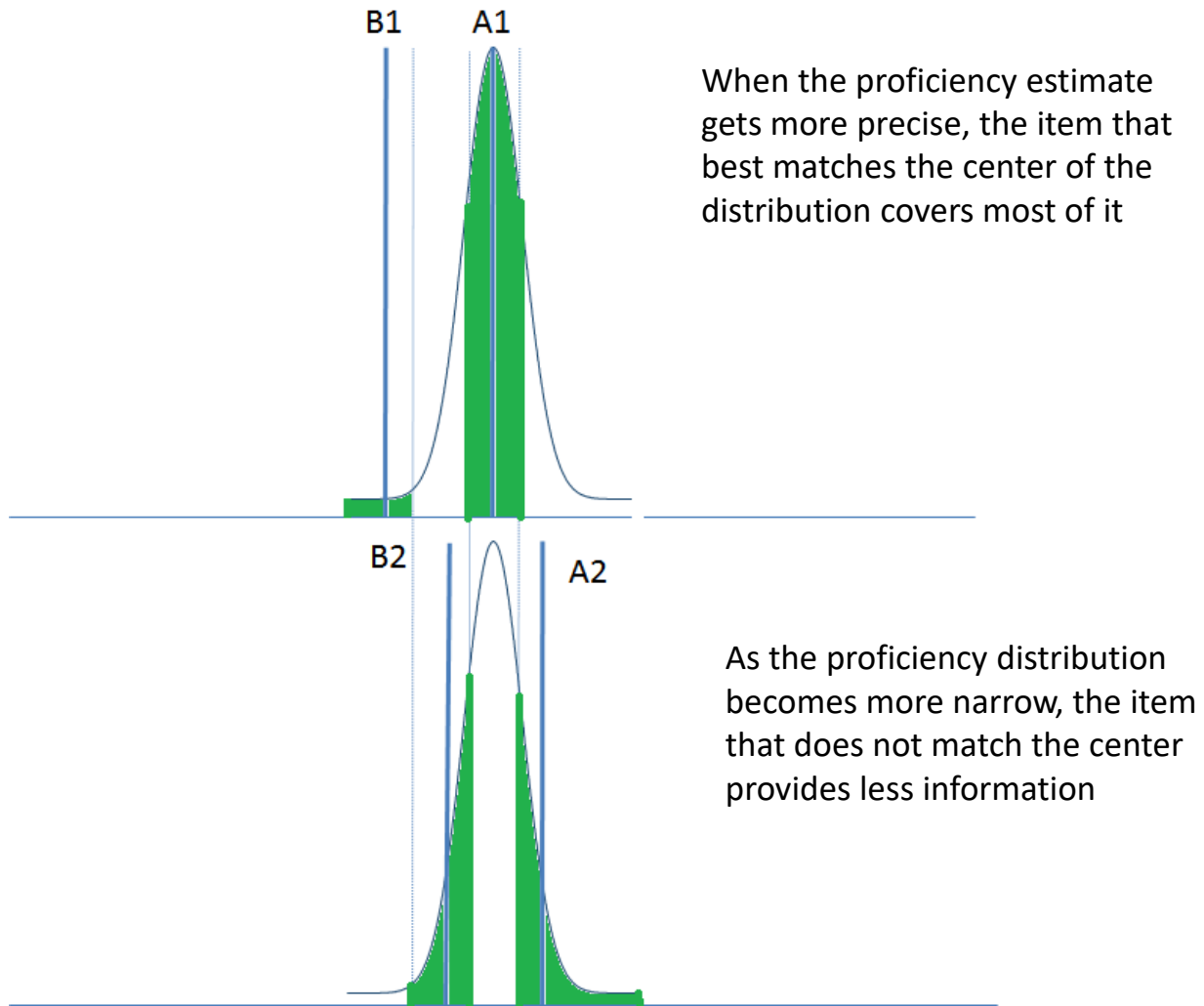


Exhibit 3: Two example items, with the shaded region showing the probability that the item maximizes information for the examinee depicted.

The approximate information value of polytomous items will be characterized as the expected information, specifically $E[I_j(\theta)|m_i, s_i] = \int \sum_{k=1}^K I_{jk}(t) p_j(k|t) \phi(t; m_i, s_i) dt$, where $I_{jk}(t)$ represents the information at t of response k to item j , $p_j(k|t)$ is the probability of response k to item j (artificially holding slope constant), given proficiency t , $\phi(\cdot)$ represents the normal probability density function, and m_i and s_i represent the mean and standard deviation of examinee i 's current estimated proficiency distribution.

We propose to use Gauss-Hermite quadrature with a small number of quadrature points (approximately five). Experiments show that we can complete this calculation for 1,000 items in fewer than 5 milliseconds, making it computationally reasonable.

As with the binary items, we propose to ignore the slope parameters to even exposure and avoid a bias toward the items with better measurement.

1.3.4 Item Group Information Value

Item groups differ from individual items in that a set of items will be selected for administration. Therefore, the goal is to maximize information across the working theta distribution. As with the polytomous items, we propose to use Gauss-Hermite quadrature to estimate the expected information of the item group.

In the case of multiple-item groups

$$E[I_g(\theta)|m_i, s_i] = \frac{1}{J_g} \int \sum_{j=1}^{J_g} I_{g(j)}(t) \phi(t; m_i, s_i) dt$$

Where $I_g(\cdot)$ is the information from item group g , $I_{g(j)}$ is the information associated with item $j \in g$, for the J_g items in set g . In the case of polytomous items, we use the expected information, as described above.

2. ENTRY AND INITIALIZATION

At startup, the system will

- create a custom item pool;
- initialize theta estimates for the overall score and each score point; and
- insert embedded field-test items.

2.1 Item Pool

At test startup the system will generate a *custom item pool*, a string of item IDs for which the student is eligible. This item pool will include all items that

- are active in the system at test startup; and
- are not flagged as “access limited” for attributes associated with this student.

The list will be stored in ascending order of ID.

2.2 Adjust Segment Length

Custom item pools run the risk of being unable to meet segment blueprint minimums. To address this special case, the algorithm will adjust the blueprint to be consistent with the custom item pool. This capability becomes necessary when an accommodated item pool systematically excludes some content.

Let

S be the set of top-level content constraints in the hierarchical set of constraints, each consisting of the tuple $(name, min, max, n)$;

\mathbf{C} be the custom item pool, each element consisting of a set of content constraints \mathbf{B} ;

f , p integers represent item shortfall and pool count, respectively; and

t be the minimum required items on the segment.

For each s in \mathbf{S} , compute n as the sum of active operational items in \mathbf{C} classified on the constraint.

$f = \text{summation over } S (\text{min} - n)$

$p = \text{summation over } S (n)$

if $t - f < p$, then $t = t - f$

2.3 Initialization of Starting Theta Estimates

The user will supply five pieces of information in the test configuration:

1. A default starting value if no other information is available
2. An indication whether prior scores on the same test should be used, if available
3. Optionally, the test ID of another test that can supply a starting value, along with
4. Slope and intercept parameters to adjust the scale of the value to transform it to the scale of the target test
5. A constant prior variance for use in calculation of working EAP scores

2.4 Insertion of Embedded Field-Test Items

Each blueprint will specify

- the number of field-test items to be administered on each test;
- the first item position into which a field-test item may be inserted; and
- the last item position into which a field-test item may be inserted.

Upon startup, select randomly from among the field-test items or item sets until the system has selected the specified number of field-test items. If the items are in sets, the sets will be administered as a complete set, and this may lead to more than the specified number of items administered.

The probability of selection will be given by $p_j = \frac{\sum_{j=1}^K K_j}{\sum_{j=1}^K a_j K_j} a_j K_j \frac{m}{N_j}$, where

p_j represents the probability of selecting the item;

m is the targeted number of field-test items;

N_j is the total number of active items in the field-test pool;

K_j is the number of items in item set j ; and

a_j is a user-supplied weight associated with each item (or item set) to adjust the relative probability of selection.

The a_j variables are included to allow for operational cases in which some items must complete field-testing sooner, or enter field-testing later. While using this parameter presents some statistical risk, not doing so poses operational risks.

For each item set, generate a uniform random number r_j on the interval $\{0,1\}$. Sort the items in ascending order by $\frac{r_j}{p_j}$. Sequentially select items, summing the number of items in the set. Stop the selection of field-test items once $FTNMin \leq m \leq FTNMax = \sum_{j=0} K_j$.

Next, each item is assigned to a position on the test. To do so, select a starting position within $f - FTMax - FTMin$ positions from $FTMin$, where $FTMax$ is the maximum allowable position for field-test items and $FTMin$ is the minimum allowable position for field-test items. $FTNMin$ and $FTNMax$ refer to the minimum and maximum number of field-test items, respectively. Distribute the items evenly within these positions.

3. ITEM SELECTION

Exhibit 3 summarizes the item selection process. If the item position has been designated for a field-test item, administer that item. Otherwise, the adaptive algorithm kicks in.

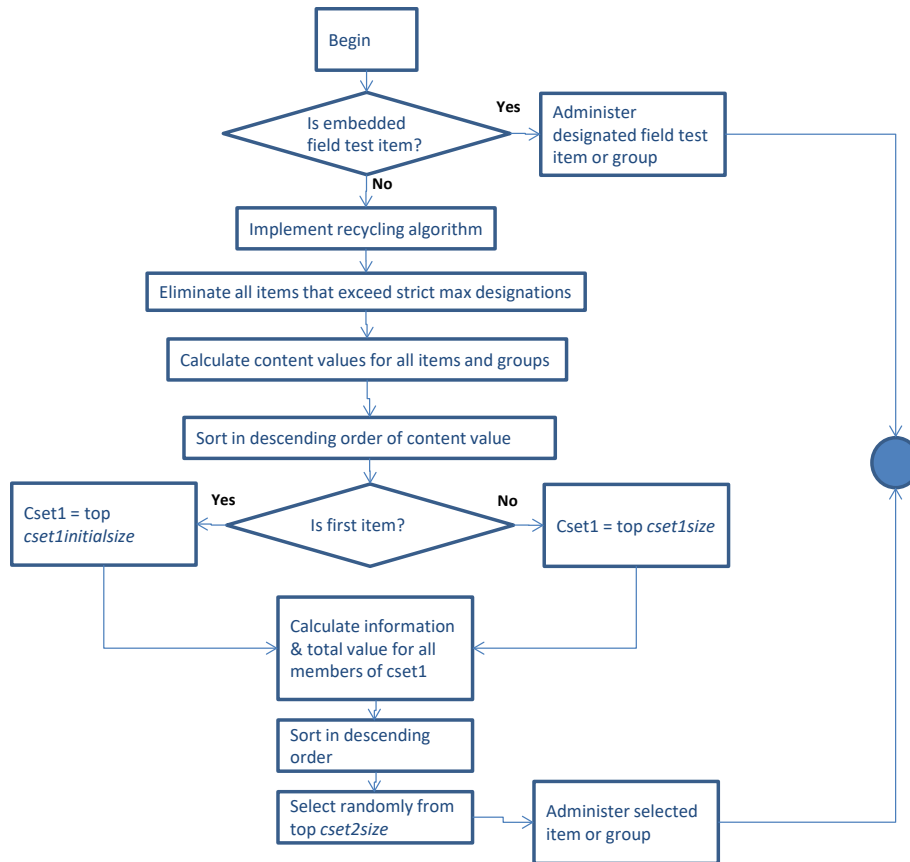


Exhibit 3: Summary of Item Selection Process

This approach is a “content first” approach designed to optimize match to blueprint. An alternative, “information first” approach, is possible. Under an information first approach, all items within a specified information range would be selected as the first set of candidates, and subsequent selection within that set would be based, in part, on content considerations. The engine is being designed so that future development could build such an algorithm using many of the calculations already available.

3.1 Trimming the Custom Item Pool

At each item selection, the active item pool is modified in four steps:

1. The custom item pool is intersected with the active item pool, resulting in a custom active item pool.
2. Items already administered on this test are removed from the custom active item pool.
3. Items that have been administered on prior tests are tentatively removed (see Section 3.2 below).
4. Items that measure content that has already exceeded a strict maximum are tentatively removed from the pool, removing entire sets containing items that meet this criterion.

3.2 Recycling Algorithm

When students are offered multiple opportunities to test, or when prior tests have been started and invalidated, students will have seen some of the items in the pool. The trimming of the item pool eliminates these items from the pool. It is possible that in such situations, the pool may no longer contain enough items to meet the blueprint.

Hence, items that have been seen on previous administrations may be returned to the pool. If there are not enough items remaining in the pool, the algorithm will recycle items (or item groups) with the required characteristic that is found in insufficient numbers. Working from the least recently administered group, items (or item groups) are reintroduced into the pool until the number of items with the required characteristics meets the minimum requirement. When item groups are recycled, the entire group is recycled rather than an individual item. Items administered on the current test are never recycled.

3.3 Adaptive Item Selection

Selection of items will follow a common logic, whether the selection is for a single item or an item group. Item selection will proceed in the following three steps:

1. Select Candidate Set 1 (*cset1*).
 - a. Calculate the content value of each item or item group.
 - b. Sort the item groups in descending order of content value.
 - c. Select the top *cset1size*, a user-supplied value that may vary by test.
2. Select Candidate Set 2 (*cset2*).
 - a. Calculate the information values for each item group in *cset1*.
 - b. Calculate the overall value of each item group in *cset1* as defined in Equation 1.
 - c. Sort *cset2* in descending order of value.
 - d. Select the top *cset2size* item groups, where *cset2size* is a user-supplied value that may vary by test.
3. Select the item or item group to be administered.
 - a. Select randomly from *cset2* with uniform probability.

Note that a “pure adaptive” test, without regard to content constraints, can be achieved by setting *cset1size* to the size of the item pool and w_2 , the weight associated meeting content constraints in Equation 1, to zero. Similarly, linear-on-the-fly tests can be constructed by setting w_0 and w_1 to zero.

3.4 Selection of the Initial Item

Selection of the initial item can affect item exposure. At the start of the test, all tests have no content already administered, so the items and item groups have the same content value for all examinees. In general, it is a good idea to spread the initial item selection over a wider range of content values. Therefore, we define an additional user-settable value, *cset1initialsize*, which is the size of Candidate Set 1 on the first item only. Similarly, we define *cset2initialsize*.

3.5 Exposure Control

This algorithm uses randomization to control exposure and offers several parameters that can be adjusted to control the tradeoff between optimal item allocation and exposure control. The primary mechanism for controlling exposure is the random selection from *CSET2*, the set of items or item groups that best meet the content and information criteria. These represent the “top *k*” items, where *k* can be set. Larger values of *k* provide more exposure control at the expense of optional selection.

In addition to this mechanism, we avoid a bias toward items with higher measurement precision by treating all items as though they measured with equal precision by ignoring variation in the slope parameter. This has the effect of randomizing over items with differing slope parameters. Without this step, it would be necessary to have other *post hoc* explicit controls to avoid the overexposure of items with higher slope parameters, an approach that could lead to different test characteristics over the course of the testing window.

4. TERMINATION

The algorithm will have configurable termination conditions. These may include

- administering a minimum number of items in each reporting category and overall;
- achieving a target level of precision on the overall test score;
- achieving a target level of precision on all reporting categories; and
- achieving a score insufficiently distant from a specified score with sufficient precision (e.g., less than two standard errors below proficient). ICCR envisions this being used in conjunction with other termination conditions to allow very high or very low achieving students to continue on to a segment that contains items from adjacent grades, but barring other students from those segments.

We will define four user-defined flags indicating whether each of these is to be considered in the termination conditions (*TermCount*, *TermOverall*, *TermReporting*, *TermTooClose*). A fifth user-supplied value will indicate whether these are taken in conjunction or if satisfaction of any one of them will suffice (*TermAnd*). Reaching the minimum number of items is always a necessary condition for termination.

In addition, two conditions will each individually and independently cause termination of the test:

1. Administering the maximum number of items specified in the blueprint
2. Having no items in the pool left to administer

A1. DEFINITIONS OF USER-SETTABLE PARAMETERS

This appendix summarizes the user-settable parameters in the adaptive algorithm.

Parameter Name	Description	Entity Referred to by Subscript Index
w_0	Priority weight associated with match to blueprint	N/A
w_1	Priority weight associated with reporting category information	N/A
w_2	Priority weight associated with overall information	N/A
q_k	Priority weight associated with a specific reporting category	reporting categories
p_r	Priority weight associated with a feature specified in the blueprint (These inputs appear as a component of the blueprint.)	features specified in the blueprint
a	Parameter of the function $h(.)$ that controls the overall information weight when the information target has not yet been hit	N/A
b	Parameter of the function $h(.)$ that controls the overall information weight after the information target has been hit	N/A
c_k	Parameter of the function $h(.)$ that controls the information weight when the information target has not yet been hit for reporting category k	reporting categories
d_k	Parameter of the function $h(.)$ that controls the information weight after the information target has been hit for reporting category k	reporting categories
cset1size	Size of candidate pool based on contribution to blueprint match	N/A
cset1initialsize	Size of candidate pool based on contribution to blueprint match for the first item or item set selected	N/A
cset2size	Size of final candidate pool from which to select randomly	N/A
cset2initialsize	Size of candidate pool based on contribution to blueprint match and information for the first item or item set selected	
t_0	Target information for the overall test	N/A
t_k	Target information for reporting categories	reporting categories
startTheta	A default starting value if no other information is available	N/A
startPrevious	An indication of whether previous scores on the same test should be used, if available	N/A
startOther	The test ID of another test that can supply a starting value, along with startOtherSlope	N/A

Parameter Name	Description	Entity Referred to by Subscript Index
startOtherSlope	Slope parameter to adjust the scale of the value to transform it to the scale of the target test	N/A
startOtherInt	Intercept parameter to adjust the scale of the value to transform it to the scale of the target test	N/A
FTMin	Minimum position in which field-test items are allowed	N/A
FTMax	Maximum position in which field-test items are allowed	N/A
FTNMin	Target minimum number of field-test items	N/A
FTNMax	Target maximum number of field-test items	N/A
a_j	Weight adjustment for individual embedded field-test items used to increase or decrease their probability of selection	field-test items
AdaptiveCut	The overall score cutscore, usually proficiency, used in consideration of <i>TermTooClose</i>	
TooCloseSEs	The number of standard errors below which the difference is considered “too close” to the adaptive cut to proceed. In general, this will signal proceeding to a final segment that contains off-grade items. Ugh.	
TermOverall	Flag indicating whether to use the overall information target as a termination criterion	N/A
TermReporting	Flag to indicate whether to use reporting category information target as a termination criterion	N/A
TermCount	Flag to indicate whether to use minimum test size as a termination condition	N/A
TermTooClose	Terminate if you are not sufficiently distant from the specified adaptive cut	
TermAnd	Flag to indicate whether the other termination conditions are to be taken separately or conjunctively	N/A

A2. SUPPORTING DATA STRUCTURES

CAI Cautions and Caveats

- Use of standard error termination conditions will likely cause inconsistencies between the blueprint content specifications and the information criteria will cause unpredictable results, likely leading to failures to meet blueprint requirements.
- The field-test positioning algorithm outlined here is very simple and will lead to deterministic placement of field-test items.

Appendix L
ICCR Item Development Plan

ICCR item development is a continuous process where items are written to enhance the already robust adaptive pool. Each development cycle begins with a bank analysis, where needs are identified as a focus for development. Standard alignment, DOK, and range of difficulty are all considered during the bank analyses. Once areas of need are identified, Item Develop Plans (IDPs) are created for each development cycle. CAI continues to develop items each year to enhance the ICCR bank, continually increasing the number of items aligned to standards, DOK, and range of difficulty. Areas of focus currently identified for West Virginia are items on the lower end of the difficulty range at every grade as well as items at the upper end for grades 7 and 8 ELA. Over the next few years of development, writing items across ranges of difficulty to enhance the pool will also be implemented. A well-rounded item pool for the adaptive assessment that spans difficulty at each standard and across all complexity levels (Depth of Knowledge), where applicable, is the goal. Having a pool of items with a greater distribution of item difficulties across all blueprint elements will help increase scoring precision.