DRC BEACON Assessments

Technical Report

2.0

Data Recognition Corporation

Developed and published by Data Recognition Corporation, 13490 Bass Lake Road, Maple Grove, MN 55311.

TABLE OF CONTENTS

LIST OF TABLES

List of Figures

Chapter 1

OVERVIEW

DRC BEACON is an interim assessment that measures student performance in English language arts and mathematics in Grades 3–8. Developed by Data Recognition Corporation (DRC), the assessment includes multiple item types that measure college- and career-ready standards.

Delivered on the DRC INSIGHT engine in a computer-adaptive test (CAT) mode, DRC BEACON's adaptive format provides students, parents, and teachers with an accurate picture of performance because the item difficulty adjusts to student ability levels. DRC BEACON can be administered up to three times a year, allowing students and teachers the opportunity to check for learning throughout the year rather than relying on a single summative test score at the end of the year.

In addition to measuring performance multiple times throughout the year, DRC BEACON reports include measurements of growth, and the interim scores can be used to predict a score range of performance on the summative test at the end of the year and/or provide a comparison to national test scores. DRC BEACON results are delivered in a dynamic interactive reporting system that allows users the opportunity to access immediate individual results, roster reports, links to college- and career-ready standards, and reports about the strengths and weaknesses of individuals and groups of students. The interactive reporting system also offers the opportunity to disaggregate, categorize, and sort data as needed.

This document provides an overview of the development of DRC BEACON and contains information about item and test development. The titles of the different chapters of the report are listed below.

- Chapter 1. Overview
- Chapter 2. Item and Test Development
- Chapter 3. Product Development Chronology
- Chapter 4. Data Analysis
- Chapter 5. Adaptive Testing
- Chapter 6. Standard Setting
- Chapter 7. Score Reporting
- Chapter 8. References

Test Configurations

DRC BEACON consists of comprehensive English language arts (ELA) and mathematics tests that can be administered in a variety of configurations throughout the year. Students may take the full ELA or mathematics assessments at one time, or they may take other test configurations called testlets that focus on specific aspects of content. For example, the ELA assessments can be administered without listening content to focus on reading and writing content only. Similarly, the assessment can be delivered using only reading items to focus exclusively on this content. The ability to select specific aspects of content for test administration means that DRC BEACON can better match the needs of

educators and their students in a variety of contexts. Table 1 provides a list of configurations available for DRC BEACON in addition to the comprehensive ELA and mathematics assessments.

Table 1. Available Test Configurations for DRC BEACON

1. English language arts Assessment a. Reading/Writing Only b. Reading Only c. Writing – Text Types and Purposes d. Writing – Conventions of Standard English e. Writing – Research f. Listening Only	2. Mathematics Assessment a. Algebra b. Numbers and Quantity c. Measurement and Data d. Geometry

A comprehensive set of scores is available when the complete ELA and mathematics tests are administered. However, when testlets are administered in lieu of the longer assessment, student performance is only based on the subset of content specified.

Test Design

DRC BEACON tests include multiple-choice items and multipoint items. All items are aligned to the college and career standards in ELA and mathematics covering grades three through eight. Each assessment is broken down into multiple reporting categories, and each item in the pool is aligned to a specific grade and reporting category. The number of items reported for each assessment and the reporting categories within the assessments are presented in Tables 2 and 3.

Table 2. Number of Items Administered within the Complete ELA Adaptive Administrations

Reporting Categories	Level/Grade	Min Items	Max Items	Passages
Full Assessment	3	56	61	6
	4	56	61	6
	5	56	61	6
	6	56	61	6
	7	56	61	6
	8	56	61	6
Reading: Key Ideas and Details	3	8	10	-
	4	8	10	-
	5	8	10	-
	6	8	10	-
	7	8	10	-
	8	8	10	-

Reporting Categories	Level/Grade	Min Items	Max Items	Passages
Reading: Integration of Knowledge & Ideas	3	8	10	-
	4	8	10	-
	5	8	10	-
	6	8	10	-
	7	8	10	-
	8	8	10	-
Reading: Vocabulary Acquisition & Use	3	8	8	-
	4	8	8	-
	5	8	8	-
	6	8	8	-
	7	8	8	-
	8	8	8	-
Writing – Text Types and Purposes	3	8	8	-
	4	8	8	-
	5	8	8	-
	6	8	8	-
	7	8	8	-
	8	8	8	-
Writing – Conventions of Standard English	3	8	8	-
	4	8	8	-
	5	8	8	-
	6	8	8	-
	7	8	8	-
	8	8	8	-
Writing – Research	3	8	8	-
	4	8	8	-
	5	8	8	-
	6	8	8	-
	7	8	8	-
	8	8	8	-
Listening	3	8	9	-
	4	8	9	-
	5	8	9	-
	6	8	9	-
	7	8	9	-
	8	8	9	-

Note. ELA administrations include 1-point and 2-point items.

Table 3. Number of Items Administered within the Complete Mathematics Adaptive Administrations

Reporting Categories	Level/Grade	Number of Items
Full Assessment	3	32
	4	32
	5	32
	6	32
	7	32
	8	32
Algebra	3	8
	4	8
	5	8
	6	8
	7	8
	8	8
Number and Quantity	3	8
	4	8
	5	8
	6	8
	7	8
	8	8
Measurement and Data	3	8
	4	8
	5	8
	6	8
	7	8
	8	8
Geometry	3	8
	4	8
	5	8
	6	8
	7	8
	8	8

Note. Mathematics administrations include 1-point items.

Various factors were considered when determining the number of operational items to administer per reporting category. The goal of DRC BEACON is to provide information on student performance in ELA and mathematics in a flexible format to meet the varying needs of educators throughout the year. The test must include enough items to provide meaningful scores with small standard errors. However, testing time must be manageable for the total test and for the various additional test configurations that DRC BEACON supports. The number of items reported for each testlet configuration and the reporting categories covered within the assessments are presented in Tables 4 through 8.

Table 4. Number of Items Administered within Reading and Writing Only Testlet

Reading and Writing Only Testlet			
Grade	Min Items	Max Items	Passages
3	48	52	4
4	48	52	4
5	48	52	4
6	48	52	4
7	48	52	4
8	48	52	4

Table 4. Number of Items Administered within Reading and Writing Only Testlet (continued)

	Reporting Categories							
	Reading: Key Ideas and Details		Reading: Integration of Knowledge & Ideas		Reading: Vocabulary Acquisition & Use	Writing – Text Types and Purposes	Writing – Conventions of Standard English	Writing – Research
Grade	Min Items	Max Items	Min Items	Max Items	Items	Items	Items	Items
3	8	10	8	10	8	8	8	8
4	8	10	8	10	8	8	8	8
5	8	10	8	10	8	8	8	8
6	8	10	8	10	8	8	8	8
7	8	10	8	10	8	8	8	8
8	8	10	8	10	8	8	8	8

Table 5. Number of Ites Administered within Reading Only Testlet

Reading Only Testlet				Reporting Categories				
				Reading: Key Ideas and Details		Reading: Integration of Knowledge & Ideas		Reading: Vocabulary Acquisition & Use
Grade	Min Items	Max Items	Passages	Min Items	Max Items	Min Items	Max Items	Items
3	24	28	4	8	10	8	10	8
4	24	28	4	8	10	8	10	8
5	24	28	4	8	10	8	10	8
6	24	28	4	8	10	8	10	8
7	24	28	4	8	10	8	10	8
8	24	28	4	8	10	8	10	8

Table 6. Number of Items Administered within Writing Testlets

Writing Testlets			
Grade	Writing – Text Types and Purposes	Writing – Conventions of Standard English	Writing – Research
	Items	Items	Items
3	10	10	10
4	10	10	10
5	10	10	10
6	10	10	10
7	10	10	10
8	10	10	10

Table 7. Number of Items Administered within Listening Only Testlet

Grade	Listening Only Testlet		
	Min Items	Max Items	Passages
3	8	10	2
4	8	10	2
5	8	10	2
6	8	10	2
7	8	10	2
8	8	10	2

Table 8. Number of Items Administered within Mathematics Testlets

Mathematics Testlets				
Grade	Algebra	Numbers and Quantity	Measurement and Data	Geometry
	Items	Items	Items	Items
3	10	10	10	10
4	10	10	10	10
5	10	10	10	10
6	10	10	10	10
7	10	10	10	10
8	10	10	10	10

Chapter 2

ITEM AND TEST DEVELOPMENT

This chapter of the *DRC BEACON Technical Report* provides a summary of the major activities involved in the development of the DRC BEACON computer-adaptive assessment. As each major activity is presented and discussed, this chapter will also highlight the role of the activity in contributing to evidence of validity for the use of the results to provide information regarding students' progression toward mastery of the DRC BEACON standards. Content-related evidence of the validity of the intended score interpretations in DRC BEACON testing is supported by the degree of correspondence or alignment between the assessment and the specifications of the standards that are assessed (i.e., what students should know and be able to do at a given grade and content area). In this chapter, content-related validity is demonstrated through DRC BEACON's consistent adherence to the assessment blueprints and through the high-quality item and test development process.

According to the most recent edition of the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014), "validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (p. 11). As stated above, essential validity evidence supporting the development of the DRC BEACON assessment is well documented through the item and test development process, including the review of the assessment items for alignment to the college- and career-ready standards that DRC BEACON measures. The information found in this chapter provides an overview of the item and test development process used for the development of the DRC BEACON computer-adaptive English language arts and mathematics assessments in grades 3–8. This chapter also includes a description of the involvement of educators, including national English language arts (ELA) and mathematics expert reviewers, throughout the development process.

DRC BEACON College- and Career-Ready Standards

DRC BEACON is a computer-adaptive assessment that measures students' progress toward mastery of a set of college- and career-ready standards that describe what students should know and be able to do at a given grade in English language arts and mathematics. It is designed to be administered to students in grades 3–8. The movement toward college- and career-ready standards in K–12 education has resulted in communicating expectations for students through the use of a set of clearly defined standards that are designed to help students, upon graduating from high school, demonstrate the preparation they need for successful progress toward college- and career-readiness. Clearly articulated college- and career-ready standards regarding what students should know and be able to do is critical for all students, whatever their pathways to graduation may be.

The college- and career-ready standards as measured by DRC BEACON are generally defined as the knowledge and skills identified within the college- and career-ready policy framework for English language arts and mathematics. This framework is based on prior consensus among leading educators, including postsecondary faculty, curriculum experts, experienced subject-matter experts, and researchers, as to what is important for teachers to teach and students to learn to be college- and career-ready. As such, the standards measured in DRC BEACON are designed to help prepare students

with the knowledge and skills they need to succeed in their future education and/or training after high school.

The standards measured by DRC BEACON are well aligned with states' standards where the foundation or framework is a set of college- and career-ready standards. This also includes alignment to the 2019 Mathematics and Reading National Assessment of Educational Progress (NAEP) Framework (National Assessment Governing Board US Department of Education, 2019). The standards measured in DRC BEACON are also well aligned with what students should know and be able to do as noted in the content curriculum standards of national consortia including the Smarter Balanced Assessment Consortium (SMARTER) and the Partnership for Assessment of Readiness for College and Careers (PARCC). As such, the evidence-based college- and career-ready standards measured by DRC BEACON utilize the research provided by a number of national organizations, including the Modern Language Association (MLA), the American Council on Education (ACE), the National Council of Teachers of English (NCTE), the National Council of Teachers of Mathematics (NCTM), the American Federation of Teachers (AFT), Achieve, and the National Education Association (NEA). In summary, the DRC BEACON assessment alignment to college- and career-ready standards includes, but is not limited to, alignment for grade-level appropriateness, depth of knowledge and cognitive complexity, and relevancy of context. A summary of the current college- and career-ready standards measured in DRC BEACON is provided in the tables below.

Table 9. English Language Arts Standards

Domain	Grade 3 Standards
Reading	Key Ideas and Details
	Craft Structure/Integration of Knowledge & Ideas
	Vocabulary Acquisition & Use
	Literary Text
	Informational Text
Writing Skills	Text Types and Purposes
	Conventions of Standard English
	Research
Listening	Listening

Domain	Grade 4 Standards
Reading	Key Ideas and Details
	Craft Structure/Integration of Knowledge & Ideas
	Vocabulary Acquisition & Use
	Literary Text
	Informational Text
Writing Skills	Text Types and Purposes
	Conventions of Standard English
	Research
Listening	Listening

Domain	Grade 5 Standards
Reading	Key Ideas and Details
	Craft Structure/Integration of Knowledge & Ideas
	Vocabulary Acquisition & Use
	Literary Text
	Informational Text
Writing Skills	Text Types and Purposes
	Conventions of Standard English
	Research
Listening	Listening

Domain	Grade 6 Standards
Reading	Key Ideas and Details
	Craft Structure/Integration of Knowledge & Ideas
	Vocabulary Acquisition & Use
	Literary Text
	Informational Text
Writing Skills	Text Types and Purposes
	Conventions of Standard English
	Research
Listening	Listening

Domain	Grade 7 Standards
Reading	Key Ideas and Details
	Craft Structure/Integration of Knowledge & Ideas
	Vocabulary Acquisition & Use
	Literary Text
	Informational Text
Writing Skills	Text Types and Purposes
	Conventions of Standard English
	Research
Listening	Listening

Domain	Grade 8 Standards
Reading	Key Ideas and Details
	Craft Structure/Integration of Knowledge & Ideas
	Vocabulary Acquisition & Use
	Literary Text
	Informational Text
Writing Skills	Text Types and Purposes
	Conventions of Standard English
	Research
Listening	Listening

Table 10. Mathematics Standards

Domain	Grade 3 Standards
Algebra	Represent and solve problems involving multiplication and division.
	Understand properties of multiplication and the relationship between multiplication and division.
	Multiply and divide within 100.
	Solve problems involving the four operations, and identify and explain patterns in arithmetic.
Number & Quantity	Use place value understanding and properties of operations to perform multi-digit arithmetic.
	Develop understanding of fractions as numbers.
Measurement & Data	Solve problems involving measurement and estimation of intervals of time, liquid volumes, and masses of objects.
	Represent and interpret data.
	Geometric measurement: understand concepts of area and relate area to multiplication and to addition.
	Geometric measurement: recognize perimeter as an attribute of plane figures and distinguish between linear and area measures.
Geometry	Reason with shapes and their attributes.

Domain	Grade 4 Standards
Algebra	Use the four operations with whole numbers to solve problems.
	Gain familiarity with factors and multiples.
	Generate and analyze patterns.
Number & Quantity	Generalize place value understanding for multi-digit whole numbers.
	Use place value understanding and properties of operations to perform multi-digit arithmetic.
	Extend understanding of fraction equivalents and ordering.
	Build fractions from unit fractions by applying and extending previous understandings of operations on whole numbers.
	Understand decimal notation for fractions, and compare decimal fractions.
Measurement & Data	Solve problems involving measurement and conversion of measurements from a larger unit to a smaller unit.
	Represent and interpret data.
	Geometric measurement: understand concepts of angle and measure angles.
Geometry	Draw and identify lines and angles, and classify shapes by properties of their lines and angles.

Domain	Grade 5 Standards
Algebra	Write and interpret numerical expressions.
	Analyze patterns and relationships.
Number & Quantity	Understand the place value system.
	Perform operations with multi-digit whole numbers and with decimals to hundredths.
	Use equivalent fractions as a strategy to add and subtract fractions.
	Apply and extend previous understandings of multiplication and division to multiply and divide fractions.
Measurement & Data	Convert like measurement units within a given measurement system.
	Represent and interpret data.
	Geometric measurement: understand concepts of volume and relate volume to multiplication and to addition.
Geometry	Graph points on the coordinate plane to solve real-world and mathematical problems.
	Classify two-dimensional figures into categories based on their properties.

Domain	Grade 6 Standards
Number & Quantity	Understand ratio concepts and use ratio reasoning to solve problems.
	Apply and extend previous understandings of multiplication and division to divide fractions by fractions.
	Compute fluently with multi-digit numbers and find common factors and multiples.
	Apply and extend previous understandings of numbers to the system of rational numbers.
Algebra	Apply and extend previous understandings of arithmetic to algebraic expressions.
	Reason about and solve one-variable equations and inequalities.
	Represent and analyze quantitative relationships between dependent and independent variables.
Geometry	Solve real-world and mathematical problems involving area, surface area, and volume.
Measurement & Data	Develop understanding of statistical variability.
	Summarize and describe distributions.

Domain	Grade 7 Standards
Number & Quantity	Analyze proportional relationships and use them to solve real-world and mathematical problems.
	Apply and extend previous understandings of operations with fractions to add, subtract, multiply, and divide rational numbers.
Algebra	Use properties of operations to generate equivalent expressions.
	Solve real-life and mathematical problems using numerical and algebraic expressions and equations.
Geometry	Draw, construct, and describe geometrical figures and describe the relationships between them.
	Solve real-life and mathematical problems involving angle measure, area, surface area, and volume.
Measurement & Data	Use random sampling to draw inferences about a population.
	Draw informal comparative inferences about two populations.
	Investigate chance processes, and develop, use, and evaluate probability models.

Domain	Grade 8 Standards
Number & Quantity	Know that there are numbers that are not rational, and approximate them by rational numbers.
Algebra	Work with radicals and integer exponents.
	Understand the connections between proportional relationships, lines, and linear equations.
	Analyze and solve linear equations and pairs of simultaneous linear equations.
	Define, evaluate, and compare functions.
	Use functions to model relationships between quantities.
Geometry	Understand congruence and similarity using physical models, transparencies, or geometry software.
	Understand and apply the Pythagorean Theorem.
	Solve real-world and mathematical problems involving volume of cylinders, cones, and spheres.
Measurement & Data	Investigate patterns of association in bivariate data.

Assessment Blueprint

AERA, APA, & NCME (2014) Standard 4.1 states the following:

> *Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended examinee population, and interpretations for intended uses. The specifications should include a rationale supporting the interpretations and uses of test results for the intended purpose(s). (p. 85)*

Adhering to AERA, APA, & NCME Standard 4.1, the key structural aspect of DRC BEACON is the assessment blueprint that specifies at a given grade and content area the target score points for each group of standards or reporting categories. The assessment blueprint was developed by a team of DRC content area experts and item and test development specialists. The experts carefully reviewed college- and career-ready standards from multiple states, consortia, and the NAEP Framework. Based on their review, a blueprint for each grade and content area was created. Each blueprint at each grade for ELA and mathematics was created to provide a road map for item development ensuring optimal content coverage and measurement of college- and career-ready standards. In general, each blueprint represented content sampling proportions that reflected the intended emphasis in instruction and mastery at each content area and grade level. Specifications for a range of items organized by standard demonstrated the desired proportions within the DRC BEACON computer-adaptive delivery constraints. In summary, the DRC BEACON assessment blueprint at each grade and content area serves to provide guidance on how the standards are measured. Provided in the tables below is the DRC BEACON blueprint for each grade and content area.

Table 11. DRC BEACON Blueprint, English Language Arts

Grade	Reporting Category*	Standards	Number of Passages	Item Ranges Per Category	Points per Item	# Min. Points
3	**Reading**		4	6–8	1–2	24
	Key Ideas and Details	RL 1, 2, 3; RI 1, 2, 3				
	Craft Structure/Integration of Knowledge & Ideas	RL 4, 5, 6, 7, 9; RI 4, 5, 6, 7, 8, 9				
	Vocabulary Acquisition & Use	L 4, 5				
	Literary Text	RL 1, 2, 3, 4, 5, 6, 7, 9; L 4, 5	2	8–10		
	Informational Text	RI 1, 2, 3, 4, 5, 6, 7, 8, 9; L 4, 5	2			
	Writing Skills		0	8	1–2	24
	Text Types & Purposes	W 1, 2, 3, 4, 5				

Grade	Reporting Category*	Standards	Number of Passages	Item Ranges Per Category	Points per Item	# Min. Points
	Conventions of Standard English	L 1, 2, 3				
	Research	W 7, 8				
	Listening	LIS 2, 3	2	8–10	1–2	8
	Total		6	56–61		56
4	**Reading**		4	6–8	1–2	24
	Key Ideas and Details	RL 1, 2, 3; RI 1, 2, 3				
	Craft Structure/Integration of Knowledge & Ideas	RL 4, 5, 6, 7, 9; RI 4, 5, 6, 7, 8, 9				
	Vocabulary Acquisition & Use	L 4, 5				
	Literary Text	RL 1, 2, 3, 4, 5, 6, 7, 9; L 4, 5	2	8–10		
	Informational Text	RI 1, 2, 3, 4, 5, 6, 7, 8, 9; L 4, 5	2			
	Writing Skills		0	8	1–2	24
	Text Types & Purposes	W 1, 2, 3, 4, 5				
	Conventions of Standard English	L 1, 2, 3				
	Research	W 7, 8, 9				
	Listening	LIS 2, 3	2	8–10	1–2	8
	Total		6	56–61		56
5	**Reading**		4	6–8	1–2	24
	Key Ideas and Details	RL 1, 2, 3; RI 1, 2, 3				
	Craft Structure/Integration of Knowledge & Ideas	RL 4, 5, 6, 7, 9; RI 4, 5, 6, 7, 8, 9				
	Vocabulary Acquisition & Use	L 4, 5				
	Literary Text	RL 1, 2, 3, 4, 5, 6, 7, 9; L 4, 5	2	8–10		
	Informational Text	RI 1, 2, 3, 4, 5, 6, 7, 8, 9; L 4, 5	2			
	Writing Skills		0	8	1–2	24

Grade	Reporting Category*	Standards	Number of Passages	Item Ranges Per Category	Points per Item	# Min. Points
	Text Types & Purposes	W 1, 2, 3, 4, 5				
	Conventions of Standard English	L 1, 2, 3				
	Research	W 7, 8, 9				
	Listening	LIS 2, 3	2	8–10	1–2	8
	Total		6	56–61		56

Grade	Reporting Category*	Standards	Number of Passages	Item Ranges Per Category	Points per Item	# Min. Points
6	**Reading**		4	6–8	1–2	24
	Key Ideas and Details	RL 1, 2, 3; RI 1, 2, 3				
	Craft Structure/Integration of Knowledge & Ideas	RL 4, 5, 6, 7, 9; RI 4, 5, 6, 7, 8, 9				
	Vocabulary Acquisition & Use	L 4, 5				
	Literary Text	RL 1, 2, 3, 4, 5, 6, 7, 9; L 4, 5	2	8–10		
	Informational Text	RI 1, 2, 3, 4, 5, 6, 7, 8, 9; L 4, 5	2			
	Writing Skills		0	8	1–2	24
	Text Types & Purposes	W 1, 2, 3, 4, 5				
	Conventions of Standard English	L 1, 2, 3				
	Research	W 7, 8, 9				
	Listening	LIS 2, 3	2	8–10	1–2	8
	Total		6	56–61		56

Grade	Reporting Category*	Standards	Number of Passages	Item Ranges Per Category	Points per Item	# Min. Points
7	**Reading**		4	6–8	1–2	24
	Key Ideas and Details	RL 1, 2, 3; RI 1, 2, 3				
	Craft Structure/Integration of Knowledge & Ideas	RL 4, 5, 6, 7, 9; RI 4, 5, 6, 7, 8, 9				
	Vocabulary Acquisition & Use	L 4, 5				
	Literary Text	RL 1, 2, 3, 4, 5, 6, 7, 9; L 4, 5	2	8–10		
	Informational Text	RI 1, 2, 3, 4, 5, 6, 7, 8, 9; L 4, 5	2			
	Writing Skills		0	8	1–2	24
	Text Types & Purposes	W 1, 2, 3, 4, 5				
	Conventions of Standard English	L 1, 2, 3				
	Research	W 7, 8, 9				
	Listening	LIS 2, 3	2	8–10	1–2	8
	Total		6	56–61		56

Grade	Reporting Category*	Standards	Number of Passages	Item Ranges Per Category	Points per Item	# Min. Points
8	**Reading**		4	6–8	1–2	24
	Key Ideas and Details	RL 1, 2, 3; RI 1, 2, 3				
	Craft Structure/Integration of Knowledge & Ideas	RL 4, 5, 6, 7, 9; RI 4, 5, 6, 7, 8, 9				
	Vocabulary Acquisition & Use	L 4, 5				
	Literary Text	RL 1, 2, 3, 4, 5, 6, 7, 9; L 4, 5	2	8–10		

Grade	Reporting Category*	Standards	Number of Passages	Item Ranges Per Category	Points per Item	# Min. Points
	Informational Text	RI 1, 2, 3, 4, 5, 6, 7, 8, 9; L 4, 5	2			
	Writing Skills		0	8	1–2	24
	Text Types & Purposes	W 1, 2, 3, 4, 5				
	Conventions of Standard English	L 1, 2, 3				
	Research	W 7, 8, 9				
	Listening	LIS 2, 3	2	8–10	1–2	8
	Total		6	56–61		56
	* Reporting for reading provided in two methods using the same passages and items					

RL = Reading Literature Text
RI = Reading Informational Text
L = Language
W = Writing
LIS = Listening

Table 12. DRC BEACON Blueprint, Mathematics

Grade	Reporting Category	Standards	# Core Items Per Category	Points per Item	# Min Points
3	**Algebra**	Operations and Algebraic Thinking	8	1	8
	Number & Quantity	Number and Operations in Base Ten Number and Operations –Fractions	8	1	8
	Measurement & Data	Measurement and Data	8	1	8
	Geometry	Geometry	8	1	8
	Total		32		32
4	**Algebra**	Operations and Algebraic Thinking	8	1	8
	Number & Quantity	Number and Operations in Base Ten Number and Operations –Fractions	8	1	8
	Measurement & Data	Measurement and Data	8	1	8
	Geometry	Geometry	8	1	8
	Total		32		32

Grade	Reporting Category	Standards	# Core Items Per Category	Points per Item	# Min Points
5	**Algebra**	Operations and Algebraic Thinking	8	1	8
	Number & Quantity	Number and Operations in Base Ten Number and Operations –Fractions	8	1	8
	Measurement & Data	Measurement and Data	8	1	8
	Geometry	Geometry	8	1	8
	Total		32		32
6	**Algebra**	Expressions and Equations	8	1	8
	Number & Quantity	Ratios and Proportional Relationships The Number System	8	1	8
	Measurement & Data	Statistics and Probability	8	1	8
	Geometry	Geometry	8	1	8
	Total		32		32
7	**Algebra**	Expressions and Equations	8	1	8
	Number & Quantity	Ratios and Proportional Relationships The Number System	8	1	8
	Measurement & Data	Statistics and Probability	8	1	8
	Geometry	Geometry	8	1	8
	Total		32		32
8	**Algebra**	Expressions and Equations Functions	8	1	8
	Number & Quantity	The Number System	8	1	8
	Measurement & Data	Statistics and Probability	8	1	8
	Geometry	Geometry	8	1	8
	Total		32		32

DRC BEACON Testlets

DRC BEACON is available in two options: in full computer-adaptive assessments in English language arts and mathematics grades 3–8 and in more focused assessments of specific aspects of content called testlets. Much like the full DRC BEACON, the testlets are adaptive and the items provided to a student will be dependent on the student's performance on previous items. Each DRC BEACON testlet will begin with items appropriate to the student's grade level and will deal easier or harder items depending on the student's performance.

The DRC BEACON testlets are designed to provide educators with flexibility. For example, some educators may choose to administer a full DRC BEACON to students at the beginning of the year to determine how well students are performing overall. They may then follow up by administering a focused DRC BEACON testlet targeted to a student's areas of need as indicated by the full DRC BEACON.

Each DRC BEACON testlet corresponds to the same reporting categories of the full DRC BEACON as follows:

For ELA, there are five DRC BEACON testlets:

- Reading (45–55 minutes),
- Writing Research (10–14 minutes),
- Writing Text Types and Purposes (10–12 minutes),
- Writing Conventions of Standard English (10–12 minutes), and
- Listening (15–20).

For mathematics, there are four DRC BEACON testlets:

- Algebra (15–20 minutes),
- Number and Quantity (15–20 minutes),
- Measurement and Data (15–20 minutes), and
- Geometry (15–20 minutes).

The blueprint for the full DRC BEACON is provided in the section of this chapter labeled "Assessment Blueprint." The blueprint for each DRC BEACON testlet is provided in the table below. The DRC BEACON testlet blueprints mirror the content of the full DRC BEACON. The DRC BEACON testlets include a subset of items from the same item pool as those included in the full DRC BEACON computer-adaptive assessments.

Table 13. Number of Items Delivered within DRC BEACON CAT Configurations – English Language Arts

	Full CAT		Testlets								
Grade	ELA w/Listening		Reading and Writing Only		Reading		Writing - Text Types and Purposes	Writing - Conventions of Standard English	Writing - Research	Listening	
	Min#	Max#	Min#	Max#	Min#	Max#	# Items	# Items	# Items	Min#	Max#
3	56	61	48	52	24	28	10	10	10	8	10
4	56	61	48	52	24	28	10	10	10	8	10
5	56	61	48	52	24	28	10	10	10	8	10
6	56	61	48	52	24	28	10	10	10	8	10
7	56	61	48	52	24	28	10	10	10	8	10
8	56	61	48	52	24	28	10	10	10	8	10

Table 14. Number of Items Delivered within DRC BEACON CAT Configurations – Mathematics

	Full CAT	Testlets			
Grade	Math	Algebra	Number and Quantity	Measurement and Data	Geometry
3	32	10	10	10	10
4	32	10	10	10	10
5	32	10	10	10	10
6	32	10	10	10	10
7	32	10	10	10	10
8	32	10	10	10	10

Item Types

Each DRC BEACON English language arts and mathematics assessment in grades 3–8 includes several types of automatically scored items. These item types are described below.

Multiple-Choice Items

Multiple-choice items are included in DRC BEACON because they are an efficient method of measuring a broad range of content and are used to assess a variety of skills, including analytical thinking. The DRC BEACON multiple-choice items require a student to select the correct answer from a group of four plausible answer options. While it is possible for a student to perform some work directly related to determining the correct answer, the student is not required to generate the content of the answer when responding to a multiple-choice item. Multiple-choice items are found in both the mathematics and English language arts assessments.

Multiple-choice items may be standalone or, for items within the reading assessment, passage based. Passage based multiple-choice items measure how well a student comprehends the overall meaning of a reading passage or makes basic inferences about the given passage or task. At times, asking students to choose a preferred answer is the best way to determine whether they have gleaned certain information from a passage.

Multi-Select Items

DRC BEACON multi-select items are a type of multiple-choice item. Multi-select items require a student to evaluate the information presented and respond by choosing two correct responses from more than four options. Multi-select items can be used to assess multiple skills and concepts.

Evidence-Based Selected-Response Items

DRC BEACON evidence-based selected-response items have two parts, and each two-part item is designed to elicit an evidence-based response from the student. For example, in English language arts, a student may be asked to read a literature passage or an informational passage and then answer an evidence-based selected-response item. In part one, which is similar to a multiple-choice item, the student may be asked to analyze the passage and choose the best answer from four response options. In part two, the student may then be asked to elicit evidence from the passage to select one or more answers based on the response to part one. This item type is only used in the DRC BEACON English language arts assessment.

Technology-Enhanced Items

The DRC BEACON technology-enhanced items are designed to elicit evidence of a broadrange of student understanding. A student interacts with the enhanced features of these computer-delivered, auto-scorable test items to show understanding of skills andconcepts. This item type uses specific enhancements, such as interactive tables and diagrams, on-screen manipulatives, and selective-response generators (e.g., drop-down lists, matching matrices), to augment the user interface. While this item type shares thesame functional structure of traditional paper-and-pencil test questions, the expansivefeatures and functions of the DRC BEACON computer-based medium allowassessments to incorporate technical enhancements into traditional elements of a test question, such as the item stem, the stimulus, the response area, or acombination of all three.

Short-Answer Items

DRC BEACON short-answer items are a type of technology-enhanced item as they require a student to enter a short numeric or algebraic response to answer a question. For the mathematics assessment, these items are designed to assess a student's ability to formulate a solution to a pure or applied mathematics problem without the assistance of response options. The short-answer items are scored using an item-specific set of scoring rules. This item type is used most often in the mathematics assessment; however, it is also used in the DRC BEACON English language arts assessment when asking a student to enter a short response such as a word or phrase.

Test Development Considerations

The major considerations in the DRC BEACON item and test development process are as follows:

- alignment to college- and career-ready standards,
- grade-level appropriateness (reading/interest level, etc.),
- freedom from issues of bias, fairness, and sensitivity,
- adherence to standards of technical quality, such as accuracy and terminology,
- adherence to grammar and item structure and style, including art and graphics,
- adherence to the Principles of Universal Design,
- content depth of knowledge and cognitive level,
- readability,
- level of complexity, and
- adherence to psychometric guidelines.

The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), the Principles of Universal Design (Thompson et al., 2002), and the DRC manual Fairness in Testing: Guidelines for Training on Bias, Fairness, and Sensitivity Issues were also used to guide the item and test development process.

Depth of Knowledge:

An important element in the development of DRC BEACON items for the computer-adaptive assessment is the process of determining the degree of alignment between the overall assessment system and the academic content standards to be measured. A methodology developed by Norman Webb (1999) was selected for use in determining the degree of alignment of the items to the college- and career-ready standards measured by DRC BEACON. The Webb model offers a comprehensive alignment method that can be applied to a wide variety of contexts. Regarding the alignment between DRC BEACON standards statements and the items included in the assessment, five categories, one of which deals with content, were used. Within each content category is a useful set of levels for evaluating the depth-of-knowledge alignment of the standards measured. According to Webb (1999), "depth-of-knowledge consistency between standards and assessment indicates alignment if what is elicited from students on the assessment is as demanding cognitively as what students are expected to know and do as stated in the standards" (pp. 7–8). The four levels of Webb's content cognitive complexity (i.e., depths of knowledge) are as follows:

- Level 1: Recall
- Level 2: Application of Skill/Concept
- Level 3: Strategic Thinking
- Level 4: Extended Thinking

The depth-of-knowledge levels were incorporated into the DRC BEACON item writing and review process, and items were coded with respect to the given level.

Passage Readability

Evaluating the readability of a passage is essentially a judgmental process by individuals familiar with the classroom context and what is linguistically appropriate at a given grade level as described in the section on reading passage selection later in this chapter. Although various readability indices were computed and reviewed in the DRC BEACON item/passage development process, it is recognized that such methods measure different aspects of readability and are often fraught with interpretive liabilities. Thus, for the passages included in the reading portion of the ELA DRC BEACON assessment, the commonly available readability formulas are not used in a rigid way. For DRC BEACON, the readability formulas are used more informally to provide several snapshots of a given passage's readability. In addition, passages are also reviewed by committees of educators who evaluate each passage for readability and grade-level appropriateness.

In the development of DRC BEACON, it is also important to note that college- and career-ready standards throughout the country consistently require students to read increasingly complex texts with greater independence and proficiency as they progress toward college- and career-readiness. As a result, DRC BEACON also includes passages that are designed to adhere to the complexity guidelines provided by the Council of Chief School Officers (CCSSO) and the English Language Arts Science State Collaborative on Assessment and Student Standards (SCASS) committee. In addition, DRC BEACON includes passages that adhere to the guidelines included in the NAEP Reading Framework. The passages within DRC BEACON include more complex forms and text structures adhering to the NAEP Framework requirements regarding the combination of literary and informational passages, including text distributions and recommended length and word counts.

Test Item Readability

In the item and test development of DRC BEACON, careful attention has also been given to the readability of the items to make certain that the assessment focus of each item did not shift based on the difficulty of reading the item. Subject areas such as mathematics contain many content-specific vocabulary terms. As a result, readability formulas are not typically used. However, wherever it is practicable and reasonable in the development of the DRC BEACON mathematics assessment, every effort has been made to keep the vocabulary one grade level below the tested grade level. As a result, for the mathematics assessment there is a conscious consideration made to ensure that each test question evaluates a student's ability to build toward mastery of the mathematics standards versus the student's reading ability. Resources used to verify the vocabulary level were the *EDL Core Vocabularies* and the *Children's Writer's Word Book.*

In addition, every test question included in DRC BEACON has been brought before several different committees of educators who review the items, stimuli, art, graphics, etc. These committees are comprised of grade-level experts in the field of ELA or mathematics education. The committee members review each question from the perspective of the students they teach, and they provide input and feedback regarding the validity of the vocabulary used and feedback regarding minimizing the level of reading required.

Items are also reviewed for bias, fairness, and sensitivity and adherence to the Principles of Universal Design. The focus of these reviews, however, is on how certain words or phrases may represent a possible source of bias or issue of fairness or sensitivity.

Bias, Fairness, and Sensitivity

At every stage of the DRC BEACON item and test development process, DRC employed procedures that are designed to ensure that items and tests adhere to the guidelines outlined in chapter 3: Fairness in Testing of the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014).

In adhering to the standards outlined in chapter 3, DRC employs a series of internal quality steps. In doing so, DRC item and test development specialists first provide specific training for item writers and reviewers on how to write, review, revise, and edit items for issues of bias, fairness, and sensitivity. Training also includes an awareness of and sensitivity to issues of cultural diversity. In addition to providing internal training in reviewing items in order to eliminate potential bias, DRC's item and test development team members in the development of DRC BEACON also provide external training to review panels of minority experts, teachers, and other stakeholders.

DRC's guidelines for bias, fairness, and sensitivity include instruction concerning how to prevent the use of language, symbols, words, phrases, and content that might be considered offensive by members of any group. Types of bias that are specifically included in the training include, but are not limited to, stereotyping, gender, regional/geographic, ethnic/cultural, socioeconomic/class, religious, and biases against an age group (ageism) or persons with disabilities.

Universal Design

As stated, the Principles of Universal Design are also incorporated throughout the DRC BEACON item and test development process to allow for the participation of the widest possible range of students taking the assessment. The following guidelines were provided in a checklist:

- Items measure what they are intended to measure.
- Items respect the diversity of the assessment population.
- Items have a clear format for text.
- Stimuli and items have clear pictures and graphics.
- Items have concise and readable text.
- Items allow changes to other forms, such as Braille, without changing meaning or difficulty.

The DRC BEACON assessment has been designed to measure knowledge and skills across the full performance continuum described in the DRC BEACON college- and career-ready standards, resulting in fairness for all students. The elements of universal design provided in the table below are addressed primarily through the physical presentation of the computer-adaptive DRC BEACON assessment. These elements include simple, clear, and intuitive instructions and procedures; maximum readability and comprehensibility; and maximum legibility.

Table 15. Elements of Universal Design

Element	Explanation
Inclusive Assessment Population	Tests designed for state, district, or school accountability and tests designed for instructional purposes, such as interim and benchmark, should have a clear goal of including every student except those in the alternate assessment. This is reflected in assessment design and field-testing procedures.
Precisely Defined Constructs	The specific constructs tested must be clearly defined so that all construct-irrelevant cognitive, sensory, emotional, and physical barriers are removed.
Accessible, Unbiased Items	Accessibility is built into items from the beginning, and bias review procedures ensure that quality is retained in all items.
Amenable to Accommodations	The test design facilitates the use of needed accommodations.
Simple, Clear, and Intuitive Instructions and Procedures	All instructions and procedures are simple, clear, and presented in understandable language.
Maximum Readability and Comprehensibility	A variety of readability and plain language guidelines are followed (e.g., sentence length and number of difficult words are kept to a minimum) to produce readable and comprehensible text.
Maximum Legibility	Characteristics that ensure easy decipherability are applied to text, tables, figures, illustrations, and response formats.

(Thompson et al., 2002)

DRC INSIGHT Adherence to the Principles of Universal Design

In adherence with the federal Individuals with Disabilities Education Act (IDEA) of 2004, DRC INSIGHT, the system used to deliver the DRC BEACON assessment to students, has been designed to be accessible to the widest possible range of students. The system is designed to ensure that appropriate accommodations are available for students with disabilities under the IDEA. INSIGHT makes available universal tools and appropriate accommodations and ensures that the assessment is accessible to students with special needs. In other words, the online system is designed to provide tools for use by all students or tools that mirror those used in instructional environments. The accommodations are appropriate and effective for meeting the individual student's need(s) to participate in taking the assessment. The accommodations are designed not to alter the construct being assessed, which allows meaningful interpretations of results and comparisons of scores for students who utilize them. The following table provides a high-level overview of the accommodations provided in the DRC BEACON INSIGHT delivery system.

Table 16. Universal Tools Available to All Students

Universal Tool	Description
Breaks/Pause	Breaks should be given based on the student's individual needs. There is no limit on the number of breaks, or the time allotted per break.
Calculators	INSIGHT provides the Desmos embedded basic and/or scientific calculator on specific items where appropriate.
Color contrast	INSIGHT allows students to adjust background or font color based on student need.
Extended time	This assessment is untimed and may be taken over several days if needed.
Flexible scheduling	A teacher can choose the time of day that is best for the student. Teachers may also stop and restart the test at any time based on the student's needs. Note: teachers may not administer any questions that had already been answered by or presented to the student.
Graphing tool	INSIGHT provides an embedded tool to graph functions on specific items where appropriate.
Highlighter	INSIGHT provides an online highlighter to be used to color text in items or passages.
Line guide	INSIGHT provides an embedded line guide that brings focus to a single line of text in items or passages.
Magnification	INSIGHT allows students to magnify the screen by 1.5 or 2 times the original size.
Masking	INSIGHT provides access to an embedded masking tool. Masking involves blocking off content that is not of immediate need or that may be distracting to the student. Students can focus their attention on a specific part of a test item by masking.
Protractor	INSIGHT provides an embedded protractor on specific items where appropriate.
Ruler	INSIGHT provides an embedded ruler on specific items where appropriate.

Universal Tool	Description
Sticky notes	Breaks should be given based on the student's individual needs. There is no limit on the number of breaks, or the time allotted per break.
Strike-through	INSIGHT provides the Desmos embedded basic and/or scientific calculator on specific items where appropriate.

Accommodations

Some students might require special accommodations to access the assessment. The DRC BEACON accommodations are provided in the table below.

Table 17. Accommodations

Accommodations	Description
American Sign Language	INSIGHT provides videos of the mathematics and ELA listening items translated into American Sign Language (ASL). This accommodation is provided in three fixed forms per grade.
Closed captioning	INSIGHT provides closed captioning for the ELA listening items. This accommodation is provided in three fixed forms per grade.
Text-to-speech (TTS)	INSIGHT provides embedded text-to-speech for all items and passages. This accommodation is available in the DRC BEACON CAT. Text-to-speech is also available in three Spanish fixed forms for mathematics.

Fixed Forms

To provide accommodations to students requiring video sign language, closed captioning, Spanish translations of the mathematics forms, and Spanish translations with text-to-speech, three unique online fixed forms have also been developed.

As stated above, DRC BEACON offers three fixed forms to provide accommodations for the following needs:

- text-to-speech in Spanish for mathematics,
- video sign language in the mathematics DRC BEACON and the English language arts DRC BEACON for the items measuring the listening content within DRC BEACON, and
- closed captioning in English language arts for the items measuring listening.

Item and Test Development Process—Detailed Description

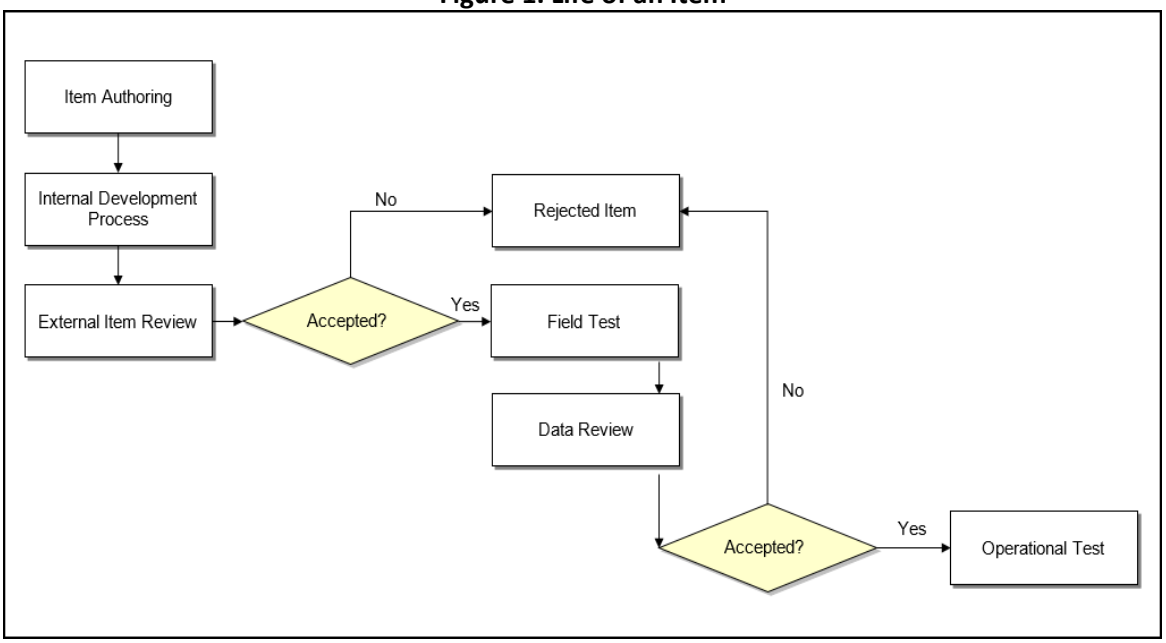
This section presents a detailed description of the DRC BEACON item and test development process, including the tasks required for the development of items, stimuli, passages, etc. As stated in this chapter, major considerations in the item development process include alignment to the DRC BEACON college- and career-ready standards, development of grade-level-appropriate items, adherence to the Principles of Universal Design in the development process, freedom from bias and sensitivity issues, style, accuracy, technical quality, etc.

The DRC BEACON items were developed and continue to be developed specifically for use in the assessment of nationally recognized college- and career-ready standards to support the measurement of these standards. After the initial development, the DRC college- and career-ready items were pilot-tested, field-tested, and operationally tested nationally. As such, the items included in DRC BEACON are aligned to college- and career-ready standards as defined in the section of this manual labeled "College- and Career-Ready Standards." In addition, the development of the items also adhered and continues to adhere to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). The development work is designed to produce a reliable and instructionally valid computer-adaptive assessment that adheres to the guidelines articulated in the AERA, APA, & NCME *Standards*. In particular, the item development process discussed in this section is in compliance with Standard 4.7, which states the following:

> *The procedures used to develop, review, and try out items and to select items from the item pool should be documented. (p. 87)*

The development of the items began in 2013 and is ongoing. The purpose of the continued item development is to replenish the computer-adaptive assessment as needed. The item and test development process used in the development of DRC BEACON items is organized to mirror the life cycle of a test item as the item moves from item authoring through review processes to operational use.

Figure 1. Life of an Item



The development of the items follows a sound method that adheres to the evidence-centered design model of development in which evidence statements are clearly noted. Emphasis is placed on developing items so that they are written to measure their respective college- and career-ready standards as required by the blueprint for a given grade and content area. In addition, items are also written to cover a range of difficulty levels and a range of subject matter.

The evidence-centered design model as stated by Joan L. Herman and Robert Linn (2015) "is a principled approach that proceeds through a series of interrelated stages to support the close correspondence between the domains about student performance that a test is designed to evaluate and the nature and quality of the assessment evidence that is used to draw inferences from student scores."

The Process for Developing Items

The sections below provide details associated with each major task in the development of the DRC BEACON items. As stated in this chapter, the first step in the process involved a careful analysis of the college- and career-ready standards as represented by states, consortia, and the NAEP Framework. The next step involved the creation of an item development plan. The development plan was primarily focused on developing items aligned to the DRC BEACON college- and career-ready standards. To ensure that the items produced were adequately distributed across groups of standards and levels of difficulty, item writers were informed of the required quantities of items to be written. Writers were then provided an item-writing template for each item. The item-writing template required writers to record additional information along with each item, such as grade level, content measured, cognitive level, and reporting category, as outlined for each college- and career-ready standard or group of standards.

Additionally, DRC carefully maintains documentation of each item writer's qualifications, including each writer's résumé and, when applicable, teaching certificate. For the development of DRC BEACON items,

further information regarding writers' qualifications is gleaned through interviews, both face-to-face and by phone.

Training Writers

Item writers have been selected and trained for the development of items in the content areas of English language arts and mathematics. DRC item and test development experts and researchers provide the training. DRC item and test development staff members are uniquely qualified to provide the training because they have received direct training in the development of college- and career-ready items by the authors and developers of the college- and career-ready framework and standards included in many state programs and consortia. In addition, DRC staff members have also been selected to write items for national consortia assessments and other college- and career-ready assessments administered in various large-scale state assessment programs. As a result, DRC staff members have a deep understanding of not only the college- and career-ready standards regarding what students should know and be able to do but also a deep knowledge and understanding of the evidence-by-design model of item development. The DRC staff's collective expertise provides the foundation for the DRC BEACON item-writing training.

The initial item-writing training for DRC BEACON took place from 2013 to 2015, and, as items are developed regularly for the purpose of future replenishment, ongoing training takes place each year. Training is typically conducted through both face-to-face meetings and virtual remote training sessions, as well as through multiple feedback conference calls. The initial training included information about how to write items to meet quality expectations, including how best to write items to adhere to the Principles of Universal Design. During the training, examples of items, stimuli, and passages were provided. The examples allowed writers to have a better understanding of what constitutes a high-quality and technically sound item to measure a given college- and career-ready standard.

After the training, writers were provided item-writing templates and other documents to help complete the assigned item-writing tasks. This information is submitted on the item-writing template and coded electronically in the DRC item-banking system, IDEAS. The DRC BEACON item bank is customized to meet the requirements of the DRC BEACON college- and career-ready standards for grades 3–8. The criteria considerations for items include, but are not limited to, the following:

- Content alignment to a given standard
- Appropriate grade level, item context, and assumed student knowledge
- Art and graphics
- Readability
- Grammar and structure for item stems and item options
- Freedom from issues of bias, sensitivity, and fairness
- Depth of knowledge and cognitive complexity
- Answer keys and scoring guidelines
- Universal design considerations (see table below)

Table 18. Universal Design Considerations

WHEN DEVELOPING ITEMS, WRITERS ASK THE FOLLOWING QUESTIONS.
Language Demand: Is the language clear, well formatted, and precise? If there is a stimulus, does the stimulus for a given set of items use correct terminology for the content area? For all students to read and use the stimulus for a given task set, they must be able to understand it. If the stimulus is formatted poorly, uses unnecessarily complex words or phrases, or uses figures or layouts that are difficult to understand, some students will not be able to answer the question(s) attached to the stimulus, or they might give incorrect answers due to these factors rather than the content being assessed.
Vocabulary and Sentence Structure: Is the vocabulary and sentence structure appropriate for each item? Vocabulary and sentence structure should not hinder students' understanding of what the item is asking students to do. In mathematics assessment, sometimes a subject-area term is used. The writer is trained to determine whether the term is needed or whether a definition can be provided.
Graphics and Displays: Are the graphics and displays of information accessible to students? Stimuli for assessment often include graphics with authentic data. These include those that are accessible to students and that model best practices in the classroom. In some cases, it may be necessary to transcribe text in a primary source to avoid possible confusion related to original type styles or vocabulary with which students may not have familiarity. It is important for students to experience and be able to analyze primary sources, but the sources must be formatted and presented so that they are accessible.

Passage Development

Passages included in DRC BEACON have been developed by English language arts and reading passage writers and published authors. Like the specialized training provided to the item writers, passage writers also receive training from DRC's college- and career-ready standards reading experts. Guidelines for passage writing include such aspects as structure, text complexity, readability, and vocabulary appropriate for the grade level. Passage writers are also trained to determine whether the reading level required by a passage is at the independent level (i.e., where the average student should be able to read 90 percent of the words in the text independently).

The DRC BEACON passage writers were initially required to write a specified number of passages for each genre. In some cases, public domain passages were acquired to address authentic works. Approval to reprint was secured from the publishers as necessary. Passages also underwent an internal review by several DRC reading specialists and national independent reviewers, who evaluated each passage's merit regarding the following criteria:

- Passages have interest value for students.
- Passages are grade appropriate in terms of text complexity, vocabulary, and language characteristics.
- Passages are free of bias, fairness, and sensitivity issues.

- Passages represent different cultures.
- Passages are from a variety of sources.
- Passages can stand the test of time.
- Passages are sufficiently rich to generate a variety of item types.
- Passages are complete with all necessary permissions documentation.
- Passages avoid dated subject matter unless a relevant historical context is provided.
- Passages do not require students to have extensive background knowledge in a certain discipline or area to understand the given passage.

As in the case of item development, passage development is ongoing, and training of passage developers for the purpose of replenishing DRC BEACON takes place each year.

Text Complexity

The DRC BEACON standards require students to read increasingly complex texts with greater independence and proficiency as they progress toward college- and career-readiness. The process used in the development or selection of DRC BEACON passages to determine text complexity involves a quantitative evaluation and a qualitative evaluation of each passage. These analyses are carefully documented for each passage. A third component, matching reader to passage text and task, is also taken into consideration during passage evaluation and review by national reading experts. Further information regarding the method by which text complexity was determined for reading passages is provided below.

Quantitative Evaluation

Evaluating the complexity of a passage involves a judgment process conducted by educators familiar with the classroom context. The process also involves understanding what is developmentally and linguistically appropriate for students at a given grade level. Readability indices, such as the Lexile, Flesch-Kincaid, Powers, and Spache measurements, are used. However, readability indices measure different aspects of readability and often result in varied interpretations. As a result, in the selection of passages included in DRC BEACON, common readability formulas have not been used in a rigid way; rather, they have been used more informally during a quantitative evaluation.

Qualitative Evaluation

DRC BEACON qualitative measures also help determine the complexity of a given passage. These include rubric-based qualitative evaluations for both literary and informational passages. The rubrics provide a comprehensive way of evaluating a range of stimulus materials that cover the literary and informational scope outlined in the DRC BEACON college- and career-ready standards. The rubrics used to determine a qualitative evaluation have been adapted from SMARTER Balanced rubrics and by the ELA Council of Chief State Officers SCASS (ELA SCASS/2012).

Quality Control

DRC BEACON items, passages, and stimuli are also reviewed by professional style editors for grammar, punctuation, and adherence to technical quality guidelines. Items, passages, and stimuli are also checked to ensure that the language is clear and consistent within and across items. The quality control checks also seek to ensure that test items follow the Principles of Universal Design, such as clear and

unambiguous items and art, limited use of shading in art, appropriate size of text in graphics, and avoidance of text on top of shading in graphics.

In addition, the quality control procedures DRC BEACON follows include not only the processes documented by DRC but also those outlined in the following resources: *Evaluating Item Quality in Large-Scale Assessment* (SCALE), the CCSSO *Criteria for Procuring and Evaluating High-Quality Assessment* (2019), and the CCSSO *Quality Control Checklist for Item Development and Test Form Construction* (SCASS/TILSA, 2005).

Internal and External Reviews

The development of high-quality items and tests depends directly on the expertise of those involved in the development effort. The items continued to be developed as required by the replenishment plan by a team of item and test development specialists who have many years of experience writing items to measure college- and career-ready standards for many state programs, including Alabama, Nebraska, Wisconsin, Michigan, Missouri, Georgia, South Carolina, Nevada, and Alaska. As an integral part of the DRC BEACON development process, content specialists, measurement experts, item and test development specialists, and professional editors review each item, passage, stimulus, etc. The team of experts evaluates each item to make sure that it measures the intended DRC BEACON college- and career-ready standard. The experts also review each item for grade-level appropriateness, and they verify that each item has the correct answer or answers for all item types included in DRC BEACON. In addition, the difficulty level, depth-of-knowledge level, graphic(s), language demand, and distractors are also evaluated. Other elements considered in this process include, but are not limited to, reviews of items, passages, and stimuli for adherence to the Principles of Universal Design, for freedom from issues of bias, and for technical quality considerations such as grammar and punctuation.

Upon completion of the internal reviews, during every development cycle, including those cycles for the purpose of replenishment of the assessment, DRC commissions a team of nationally known subject-area experts who review items for alignment, content, bias, adherence to the Principles of Universal Design, technical quality, etc. For the initial 2013 DRC BEACON development, 32 national reviewers provided reviews and feedback on the items, passages, and stimuli. These external reviewers had a broad range of experience in the educational field. All the reviewers had bachelor-level, master-level, or doctoral-level degrees and teaching experience in their specific areas of expertise. Information regarding the expert reviewers can be found in the section below. Overall, the knowledge and educational experience of the item and passage writers and the national reviewers met the requirements of the following AERA, APA, & NCME (2014) Standards:

> **Standard 3.1** *Those responsible for test development, revision, and administration should design all steps of the testing process to promote valid score interpretations for intended score uses for the widest possible range of individuals and relevant subgroups in the intended population*. *(p. 63)*

> **Standard 3.2** *Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests' being affected by construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics*. *(p. 64)*

Reviewers

The ten DRC BEACON English language arts reviewers have vast coursework backgrounds in the fields of English language arts, reading, and curriculum; English as a second language; talented and gifted programs; elementary, middle, and secondary education; postsecondary education; and applied linguistics. They represent all levels in the field of teaching, from kindergarten through collegiate as well as Title I, Chapter I, and special education. Many provide professional development to preservice teachers and current teachers, including work in the implementation of college- and career-ready standards and instructional strategies. Reviewers are located throughout the country (i.e., Alabama, California, Florida, Illinois, Iowa, Kansas, Michigan, North Carolina, Pennsylvania, and Texas). They represent many backgrounds and provide a national understanding and perspective of college- and career-readiness.

The ten DRC BEACON mathematics reviewers are current or former teachers who have a range of experiences and expertise in the field of mathematics education. For example, all reviewers have experience teaching in K–12 classrooms and had extensive knowledge of what students should know and be able to do regarding college- and career-ready standards. Many reviewers have also had teaching experience at the undergraduate and/or graduate level to prepare future teachers to understand what is required instructionally to help students progress toward mastery of the standards. Others have backgrounds in assessment, including working with diverse populations of students and those with special needs. In addition, all the reviewers DRC selects for each development cycle have extensive experience with college- and career-readiness. Much like the English language arts reviewers, mathematics reviewers also reside in various parts of the country, including West Virginia, Alabama, Minnesota, Wisconsin, and Georgia.

The ten DRC BEACON reviewers of bias, fairness, and sensitivity also have a vast array of experience in education, which provides them with diverse perspectives. All reviewers selected are those experienced in the review of passages and items in English language arts and mathematics for bias, fairness, and sensitivity and for adherence to the Principles of Universal Design. Their perspectives and experiences include, for example, knowledge of special populations, such as those with limited English language proficiency. Their professional backgrounds also include classroom teacher (e.g., regular education, special education, and gifted/talented education), curriculum specialist, content area instructional specialist, test development editor, university professor, adjunct professor, disability rights advocate, and superintendent. Additionally, the reviewers reside in several areas (e.g., Arkansas, California, District of Columbia, Illinois, Pennsylvania, Wisconsin), providing a national perspective. Many of the reviewers are published authors with publications in the field of education.

Table 19. ELA External Reviewers of the College- and Career-Readiness Item Bank

Reviewer	# of Years of Experience	Highest Degree
Reviewer 1	11	MA, Curriculum and Instruction
Reviewer 2	9	PhD, American Literature
Reviewer 3	5	EdD, Reading & Language Arts
Reviewer 4	21	EdD, Curriculum and Instruction
Reviewer 5	34	PhD, Educational Research Methodology
Reviewer 6	19	MA, Applied Linguistics
Reviewer 7	23	PhD, English Education
Reviewer 8	29	MA, Education; emphasis in Reading
Reviewer 9	26	MA, English
Reviewer 10	25	PhD, Curriculum and Instruction

Table 20. Mathematics External Reviewers of the College- and Career-Readiness Item Bank

Reviewer	# of Years of Experience	Highest Degree
Reviewer 1	19	EdD, Education Leadership
Reviewer 2	43	EdM, Secondary Mathematics Education
Reviewer 3	36	PhD, Mathematics Education
Reviewer 4	30	MS, Mathematics Education
Reviewer 5	19	BS, Mathematics, Licensure Secondary Mathematics Education
Reviewer 6	44	PhD, Mathematics Education
Reviewer 7	2	EdD, Curriculum and Instruction
Reviewer 8	8	EdM, Mathematics Education
Reviewer 9	6	EdM, Mathematics Education
Reviewer 10	6	EdM, Mathematics Education

Table 21. Bias and Sensitivity External Reviewers of the College- and Career-Readiness Item Bank

Reviewer	# of Years of Experience	Highest Degree
Reviewer 1	46	BS, Elementary Education
Reviewer 2	20	PhD, English; emphasis in Multicultural Pedagogy
Reviewer 3	20	PhD, Education
Reviewer 4	34	PhD, Rhetoric and Linguistics
Reviewer 5	44	MA, Curriculum and Supervision
Reviewer 6	44	BA, History and Secondary Education; emphasis in Gifted Education
Reviewer 7	26	MA, Education; emphasis in Special Education
Reviewer 8	34	MA, Education; emphasis in Special Education
Reviewer 9	53	MA, Education; MS, Social Studies
Reviewer 10	27	BA, Spanish

Chapter 3

PRODUCT DEVELOPMENT CHRONOLOGY

The development of an assessment occurs in multiple phases. The initial planning and design phase consisted of an extensive review of the curriculum. Content specifications were then developed, and test blueprints were created to measure the content specifications. Items were then written to fill out the test blueprints, and plans for item piloting were developed.

Pilot testing was conducted in multiple states (Alabama, Alaska, California, Kentucky, Louisiana, Michigan, Minnesota, Nebraska, Ohio, Oregon, and Texas). The major purposes of the pilot tests were to administer items to obtain initial item classical statistics, to evaluate the protocols for the test administration and computer delivery system (i.e., technology infrastructure), and to implement targeted test accommodations and elements of Universal Design. Another important goal of the pilot test was to pilot a variety of new technology-enhanced item types to determine the best use of the item type when assessing a given DRC BEACON college- and career-ready standard. In total, over 5,000 items were pilot tested, and the items were administered using a fully randomized design for each subject and grade.

After pilot data were analyzed and reviewed, a final set of items was selected to be the basis of the DRC BEACON item banks and field-tested. Field-testing took place in 2016 and 2017 in several states. Selection of the items to be field-tested and construction of the various test forms used in states where field-testing took place was a collaborative effort between DRC's item and test development specialists and the DRC psychometric services team. This selection process involved a series of steps to determine the technical quality of each item and included a confirmation of the alignment of each item to a given college- and career-ready standard.

Field-testing involved a specific set of guidelines related to the selection of items for embedding other summative assessments. DRC psychometricians examined the statistical quality of the items based on the pilot testing, paying specific attention to *p*-value and discrimination targets and associated distracter analyses. In addition, test development staff reviewed field test configurations to ensure that there would be no potential issues related to developmental appropriateness, item cueing, or redundant content. Field-testing involved large sample sizes and a series of extensive psychometric analyses to calibrate and express item parameters on a vertical scale of measurement used to support scoring and reporting for DRC BEACON.

A brief description of the design chronology for the DRC BEACON assessments follows.

Test Planning and Design (2014–2015)

- Conducted curriculum review and developed content specifications
- Designed test blueprints and planned test configurations
- Planned tryout and field test research

Item Development (2014)

- BEACON 1.0 items developed

Item Piloting (2015)

- Collected pilot test data
- Conducted classical item analyses
- Evaluated technology-enhanced scoring rules
- Validated targeted test accommodations and universal design

Field-Testing and Scale Development (2016–2017)

- Collected statistical data
- Calibrated items using item response theory
- Established the vertical scale
- Conducted bias analyses
- Built item pools
- Built and evaluated Initial CAT Configurations

BEACON 1.0 Released

- Initial configurations available (2018–2019)
- ELA and Mathematics Assessments
- Reading Only testlet
- Writing Only testlet

DRC BEACON 2.0 Availability

- Updated CAT configurations available (2020–2021)
- ELA and Mathematics Assessments
- Six ELA testlets
- Four Mathematics testlets

Chapter 4

DATA ANALYSIS

Calibration and Scaling

The analysis plan for DRC BEACON consisted of three major types of analysis: classical item analysis, item response theory calibration, and vertical scaling. The analysis components were carefully implemented to produce the psychometric infrastructure required to support adaptive test delivery of the DRC BEACON assessments. Each type of analysis is discussed briefly below, and key elements of the psychometric work required for the development of DRC BEACON are discussed in this section.

Item Analyses

The items developed to populate the DRC BEACON item bank were field-tested during the spring 2017–2018 test administrations. All items were evaluated using a comprehensive set of item analysis statistics based on the field test data. Classical item analysis provided item functioning statistics for multiple-choice items and analogous information for multipoint items, including *p*-values, item-total correlations, and associated distractor analyses. Differential item functioning (DIF) statistics were also estimated to ensure that the items were functioning equivalently for different gender and ethnic subgroups. Definitions of relevant statistical terms are provided below.

p-value: The *p*-value is a measure of item difficulty. For a multiple-choice item, the *p*-value is calculated by taking the number of students who correctly responded to an item and dividing by the total number of students who attempted the item. The value is reported as a proportion. For a constructed-response item, the *p*-value is calculated by taking the average score for the item and dividing by the maximum points possible and is also reported as a proportion.

Item-Total Correlation: An item-total correlation is the correlation between an item and the total test score, where the item score is included in the total score. It indicates how well an item differentiates between low- and high-achieving students.

Distractor Analysis: Distractor analyses refer to the evaluation of the relative magnitude of various statistics for the different scorable components of an item. For example, a multiple-choice item with an incorrect key would likely exhibit one or more of the following characteristics:

- Proportion correct (*p*-value) is low;
- Percent of students selecting any distractor is high;
- Point-biserial correlation for the key is low; or
- Point-biserial correlation for a distractor is high.

Differential Item Functioning: DIF statistics are used to quantify differences in item performance between two groups after controlling for examinees' overall performance level. The DIF statistics indicate the degree to which members of a particular subgroup perform better or worse than expected on each item as compared to a reference group. It should be noted, though, that all items included on the DRC BEACON assessment have been thoroughly reviewed for content and bias to ensure that they

do not tap knowledge or specific ability irrelevant to the construct the test intends to measure. Therefore, a DIF flag does not necessarily indicate that an item is biased; rather, a DIF flag indicates that the item functions differently for equally able members of different groups (Camilli & Shepard, 1994). Items are not automatically omitted from operational scoring if they are flagged for DIF. However, items exhibiting large DIF are typically avoided for use in operational assessments. Given that DRC BEACON was built using item response theory methodology, DIF analyses using an approach developed by Linn and Harnisch (1981) were implemented for gender and ethnic groups.

Statistical criteria were set to flag items within the DRC BEACON item analyses for possible defects in quality related to content, bias, or accessibility. Criteria that triggered item review are in Table 22. Items with no statistical flags were eligible for potential use in the operational pools.

Table 22. Item Flagging Criteria

Flag	Definition
1	High difficulty (*p*-value less than 0.10)
2	Items with proportionally more high-proficient students selecting a distractor over the key
3	Low difficulty (*p*-value greater than 0.95)
4	Items with positive item total correlations for distractors
5	Low item-total correlation (less than 0.20)
6	Items with C-Level DIF for any subgroup

Tables 23 and 24 summarize the number of items that were flagged for DIF for each group. The analyses were conducted by grade. ELA had 0–29 items flagged for large DIF, while mathematics had 0–5 items, positive or negative, across male, female, Caucasian, Asian, Black, Hispanic, American Indian, and Pacific Islander subgroups.

Table 23. Differential Item Functioning Flagged Items: English Language Arts

Grade	DIF Category	Male	Female	Caucasian	Asian	Black	Hispanic	American Indian/ Alaska Native	Pacific Islander
3	-	16	16	11	8	6	15	8	0
	+	17	15	16	10	12	12	7	0
4	-	26	28	19	15	18	29	13	0
	+	28	27	20	13	20	22	16	0
5	-	0	0	0	0	5	2	0	0
	+	2	2	1	1	2	0	2	1

Grade	DIF Category	Male	Female	Caucasian	Asian	Black	Hispanic	American Indian/ Alaska Native	Pacific Islander
6	-	2	4	1	1	5	1	0	2
	+	0	3	0	2	2	4	0	2
7	-	3	1	1	2	4	2	0	1
	+	1	1	0	1	2	3	0	1
8	-	0	1	1	2	2	1	1	1
	+	0	0	0	1	0	1	0	0

Table 24. Differential Item Functioning Flagged Items: Mathematics

Grade	DIF Category	Male	Female	Caucasian	Asian	Black	Hispanic	American Indian/ Alaska Native	Pacific Islander
3	-	2	2	2	5	2	2	1	1
	+	0	0	0	0	0	0	1	1
4	-	4	3	2	2	2	2	3	1
	+	0	0	0	0	0	0	0	1
5	-	2	2	2	2	5	1	0	0
	+	0	1	0	1	2	0	0	0
6	-	2	3	1	1	0	2	1	2
	+	1	0	0	1	0	1	1	0
7	-	0	0	0	2	0	0	0	0
	+	1	3	1	2	3	3	1	2
8	-	0	0	0	2	0	1	0	0
	+	0	0	0	2	0	0	0	0

Item Response Theory Calibration

To ensure an accurate description of test performance, Item Response Theory (IRT) models were employed throughout the development of DRC BEACON including the item calibration and the construction of the vertical scale used to support student scoring and reporting.

The item response theory (IRT) model used in DRC BEACON is the Generalized Partial Credit Model (GPCM) (Muraki, 1992):

$$P_{im}(\theta_j) = \frac{exp\left[\sum_{k=0}^{m} Da_i(\theta_j - b_{ik})\right]}{\sum_{v=0}^{M_i-1} exp\left[\sum_{k=0}^{v} Da_i(\theta_j - b_{ik})\right]},$$

where $\sum_{k=0}^{0} Da_i(\theta_j - b_{ik}) \equiv 0$; $P_{im}(\theta_j)$ is the probability of an examinee with ability θ_j getting score *m* on item *i*; M_i is the number of score categories of item *i* with possible scores as consecutive integers from 0 to $M_i - 1$; D is the scaling constant 1.7; a_i is the discrimination parameter of item *i*; and b_{ik} is the location parameter or threshold of category *k*. The GPCM is equivalent to the Two-Parameter Logistic (2PL) Model (Birnbaum, 1968) when the item is scored dichotomously. The 2PL model is shown below:

$$P_i(\theta_j) = \frac{1}{1+exp\left[-Da_i(\theta_j - b_i)\right]},$$

where $P_i(\theta_j)$ is the probability of an examinee with ability θ_j answering item *i* correctly; D is the scaling constant 1.7; and a_i and b_i are the discrimination and difficulty parameters of item *i*.

Note that the DRC BEACON item pool consisted of selected-response (SR), multi-select (MS), short-answer (SA), evidence-based selected-response (EBSR), and technology-enhanced (TE) items. Items with a maximum item score of 1 point used the 2PL Model, and items with a maximum item score of more than 1 point were calibrated with the GPCM. However, as described above, the GPCM is equivalent to the 2PL Model when an item's maximum score is 1 point.

The IRT models were implemented using the PARDUX software (DRC, 2015). PARDUX estimates parameters simultaneously for dichotomous and multipoint items using marginal maximum likelihood procedures implemented with the EM algorithm (Bock and Aitkin, 1981; Thissen, 1982). Extensive simulation studies and comparisons between PARDUX and other programs (i.e., WINSTEPS, MULTILOG, PARSCALE) have shown that PARDUX provides the same or more precise parameter and ability estimates (Fitzpatrick, 1991; Fitzpatrick & Julian, 1996).

After the initial item calibrations, goodness-of-fit statistics were computed for each item to examine how closely the item's data conform to the item response models. The Q1 index described by Yen (1981) was used to measure data-model fit. Poor-fitting items are potentially revised and field-tested again.

Sample Description

The data were obtained from a large, convenient sample of students in public and private schools across the country and were structured to be heterogeneous. A subset of items were administered at multiple grades to facilitate the development of vertical scales. The sample sizes for the calibration and scaling of mathematics and ELA items are reported in Table 25 below. Information about the sample composition in terms of gender and ethnicity is provided in Table 26.

Table 25. Sample Size

Grade	Sample Size	
	Math	ELA
3	81,157	68,728
4	93,082	79,901
5	84,320	72,590
6	84,025	71,977
7	86,158	75,458
8	67,851	61,195

Table 26. Number of Examinees in DRC BEACON Sample by Gender and Ethnicity/Race

	Female		Male		Missing		Total	
Ethnicity/Race	Frequency	Percent	Frequency	Percent	Frequency	Percent	Frequency	Percent
African American or Black	21,151	49.4	21,687	50.6	.	.	42,838	10.00
American Indian or Alaska Native	53,366	48.7	56,087	51.2	23	0.0	109,476	25.5
Asian	4,414	52.4	4,003	47.6	.	.	8,417	2.0
Native Hawaiian or Other Pacific Islander	300	56.2	234	43.8	.	.	534	0.1
Hispanic or Latino	18,417	49.0	19,188	51.0	4	0.0	37,609	8.7
Caucasian	104,493	48.5	111,176	51.5	.	.	215,669	50.2
Mixed/2 or more	7,262	49.7	7,355	50.3	.	.	14,617	3.4
Missing	232	33.7	297	43.1	160	23.2	689	0.2
Total	209,635	.	220,027	.	187	.	429,849	.
Percentage	.	48.8	.	51.2	.	0.0	.	.

Vertical Scaling

The DRC BEACON vertical scales were constructed using user data from several states, with a total sample of 61,195–93,082 students per grade and content area. The vertical scales were constructed across grades 3 through 8 for both English language arts and mathematics.

A multi-group concurrent calibration was chosen for the vertical scale linking based on a common item anchor design, with some samples taking both an on-grade and below-grade linking form. This method estimated the mean and standard deviation of the ability distribution for each grade group along with the item parameters for all items across all levels. DRC's proprietary software PARDUX (DRC, 2015) was used. It utilizes a Marginal Maximum Likelihood procedure for item parameter estimation and a Maximum Likelihood procedure for person parameter estimation. Fifth grade students were assumed to have a standard normal distribution with a mean of zero and a standard deviation of one in order for the model to be identified. A linear transformation was then applied using a mean of 500 and a standard deviation of 70 for ELA and 90 for mathematics to put the parameters onto the new DRC BEACON scale score metric.

A comparison of three vertical scaling methods on the same data set (Karkee et al., 2006) and vertical scaling in common item design (Karkee et al., 2003) showed that the multi-group concurrent method provides similar or, in many circumstances, better item parameter estimates and scaling results in terms of standard errors of measurement, level-to-level growth, level-to-level variability, and separation of scores across grade levels.

Figures 2 and 3 show the banked item location estimates for the operational item pools of each grade for the ELA and mathematics vertical scales. The plots show the vertical articulation of the items that support the DRC BEACON adaptive test configurations.

Figure 2. Banked Item Location Estimates for ELA

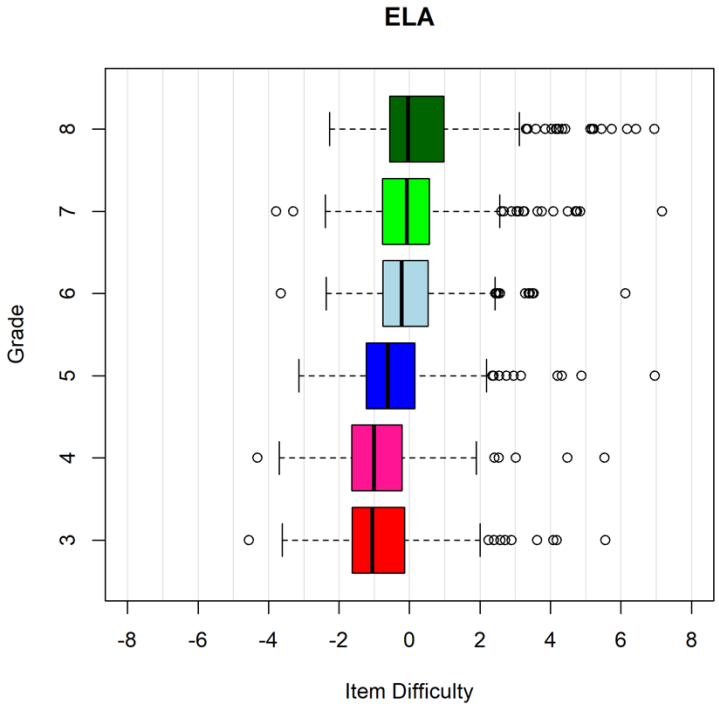
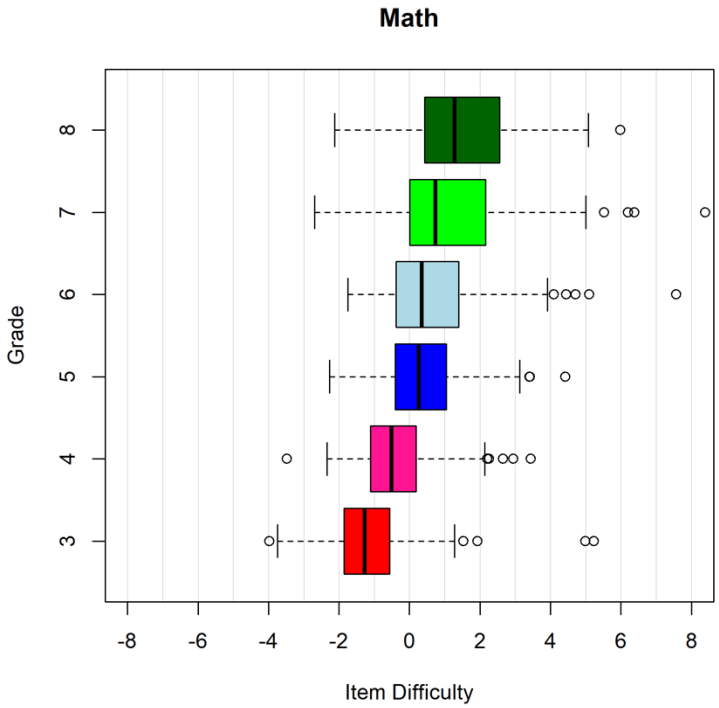


Figure 3. Banked Item Location Estimates for Mathematics



Ability Estimates and Standard Error of Measurement (SEM)

When assessments are calibrated and scaled using IRT, student ability estimates are reported on the same scale of measurement that is used to express the item parameters. DRC BEACON ability estimates and associated SEM are calculated via the maximum likelihood estimation (MLE) for the total test score and for each reporting category score. As described in the Item Response Theory Calibration section, items in the item bank are calibrated based on the GPCM.

For a general MLE, the likelihood combines both single and multipoint items as shown below:

$$L(\theta_j|U) = \left[\prod_{i=1}^{n} P_i(\theta_j)^{u_i} Q_i(\theta_j)^{1-u_i}\right]\left[\prod_{i=n+1}^{N}\prod_{m=0}^{M_i-1} P_{im}(\theta_j)^{u_{im}}\right],$$

where $Q_i(\theta_j)$ is $1 - P_i(\theta_j)$, and the response matrix *U*, shown below, contains the response of dichotomous items (single point items):

$$u_i = \begin{cases} 1, & if\ correct, \\ 0, & otherwise, \end{cases}$$

for *i* = 1,…, *n*, and the responses of polytomous items (multipoint items):

$$u_{im} = \begin{cases} 1, & if\ scored\ m, \\ 0, & otherwise, \end{cases}$$

for *i* = *n*+1, …, *N* and m = 0, 1, …, $M_i - 1$.

The Newton-Raphson equation for estimating theta at iteration *t* is given as shown below:

$$\hat{\theta}_t = \hat{\theta}_{t-1} + \frac{L_1' + L_2'}{ABS(L_1'' + L_2'')}.$$

ABS stands for the absolute value. L_1' and L_1'' are the first and second derivative of the likelihood function of dichotomously scored items,

$$L_1' = \sum_{i=1}^{n} D a_i (u_i - p_i),$$

and

$$L_1'' = \sum_{i=1}^{n} \frac{D^2 a_i^2 (-p_i^2)(1 - p_i)}{p_i},$$

where u_i is the score a student gets from a dichotomously scored item, with possible values of 1 or 0.

L_2' and L_2'' are the first and second derivative of the likelihood function of polytomously scored items,

$$L_2' = \sum_{i=n+1}^{N} Da_i \sum_{m=0}^{M_i-1} u_{im} \left[m - \sum_{m=0}^{M_i-1} mP_{im}(\theta_j) \right]$$

and

$$L_2'' = -\sum_{i=n+1}^{N} D^2 a_i^2 \left[\sum_{m=0}^{M_i-1} m^2 P_{im}(\theta_j) - \left[\sum_{m=0}^{M_i-1} mP_{im}(\theta_j) \right]^2 \right]$$

where $u_{im}$is the value 1 or 0.

For each ability (i.e., theta) estimate, the corresponding SEM is calculated. SEM is the inverse of the square root of the test information function (TIF), which is the sum of the item information functions (IIF). The IIF for dichotomously and polytomously scored items can be calculated by using the following equations respectively:

$$IIF_i = D^2 {a_i}^2 (1 - P_i) P_i$$

and

$$IIF_i = D^2 a_i^2 \left[\sum_{m=0}^{M_i-1} m^2 P_{im} - \left[\sum_{m=0}^{M_i-1} mP_{im} \right]^2 \right]$$

It is important to note that the majority of the IRT calibration and scaling work is implemented on the theta scale and that this scale is linearly transformed for reporting purposes. In addition, the MLE procedures employed in the IRT procedures cannot produce scale score estimates for students with perfect scores or scores of zero. Also, while MLEs are available for students with extreme scores other than perfect or zero, these estimates occasionally have standard errors of measurement that are very large, and differences between these extreme values have little meaning. Therefore, scores are established for these students based on a rational but necessarily non-maximum likelihood procedure. These values are set separately by levels and are called the lowest obtainable scale score (LOSS) and the highest obtainable score (HOSS). The scaling constants and the LOSS and HOSS values used to support DRC BEACON are reported in Table 27.

Table 27. Scale Transformation Constants and LOSS/HOSS

Subject	Grade	Slope A	Intercept B	LOSS	HOSS
ELA	3	70	500	160	800
	4	70	500	180	820
	5	70	500	200	840
	6	70	500	220	860
	7	70	500	240	880
	8	70	500	260	900
Math	3	90	500	160	800
	4	90	500	180	820
	5	90	500	200	840
	6	90	500	220	860
	7	90	500	240	880
	8	90	500	260	900

Chapter 5

ADAPTIVE TESTING

The DRC BEACON assessments are delivered using computer-adaptive testing (CAT) software that is designed to be more efficient in that the items or sets of items that are selected for a student as they progress through the test are not too difficult or too easy. Relative to fixed form testing, adaptive testing can save time without sacrificing the quality of the information obtained. Moreover, the testing experience for every student is optimized in such a manner that all students are engaged consistently throughout each test administration. The adaptive testing software that supports DRC BEACON runs in the background as a student completes an assessment.

In an adaptive test like DRC BEACON, a student will initially be administered a few items of average difficulty. Then the adaptive software will select subsequent items that meet the test blueprint specifications while concurrently matching the subsequent item difficulty to the student's performance as they move through the assessment. The specifications include the content to be covered by the assessment, the number of items to be administered, the number of score points, and the degree to which off-grade content can be used throughout the test. The test ends once the requirements specified in the assessment configurations have been met and there is enough information to provide test scores as intended. Ability estimates for total scores and subtest scores, along with the associated standard error of measurement, are then reported for the students.

This chapter details the design and specifications of the CAT algorithm used to deliver DRC BEACON. The aspects of ELA and mathematics content that are covered within test administrations, called reporting categories, and the number of items to be administered per reporting category are provided. Additional features including the entry point into the adaptive testing, the item selection criteria, test navigation, and test termination are covered within this section as well.

Computer-Adaptive Test Algorithm

Computer-adaptive testing uses an algorithm to ensure that each student is administered a test that covers the required content and presents items that match the student's ability. The algorithm operates in the background, selecting the next item to be administered by considering what elements of the test blueprint need to be covered and the student's performance on all prior items. Moving throughout the test administration, a student's prior performance on items determines the ease or difficulty of the remaining content, such that in the end, the information available regarding the student's ability has been maximized while ensuring the student has been administered an assessment that was optimally constructed for the student's ability. The algorithm is essentially a set of rules that govern the features of the adaptive administration including the entry point into the CAT assessments, the item selection criteria, the test navigations, and test terminations.

Entry Point

Adaptive tests are designed to administer items targeted to each student based on their performance on prior content. However, student performance is often an unknown at the beginning of a test. With

no prior information available about student performance, the starting point for DRC BEACON tests is a small locator section in which a small number of items of average difficulty from several reporting categories are administered. The student's grade is considered when the algorithm determines what constitutes the average difficulty of the items within reporting categories. For example, if a 6th grader is taking a DRC BEACON test for the first time, the algorithm will identify an item of average difficulty in each reporting category that is aligned with 6th grade content standards. The sequence of the reporting categories in which these initial items are administered to students is randomly determined. It is important to note that passage-based items in DRC BEACON are not initially administered to students.

The CAT algorithm includes a randomization component when selecting items to control item exposure. Rather than identifying a single item that is equal to the average difficulty, one item is randomly selected from among a set of items that are near the targeted item difficulty. This is especially important at the beginning of the DRC BEACON when no prior information is available. Randomization of items and sequence of reporting categories administered initially ensures that students will not see the same set of items in the same order even when all students are assigned items of average difficulty.

If a student has previously taken a DRC BEACON assessment, the prior scores are used to give the algorithm a head start. In that case, the first item in each reporting category is selected to match the characteristics of the prior score information. For example, if a student previously took a full DRC BEACON mathematics assessment and performed very high in Algebra relative to the other reporting categories, then the first Algebra item selected in a subsequent administration will be more difficult than the items selected for the other reporting categories.

Item Selection Criteria

Once the initial set of items has been administered, the CAT algorithm is designed to administer items targeted for the individual student based on performance. In selecting items to be administered to a student throughout the administration, the CAT algorithm uses item response theory ability estimates from the current test session and considers several factors including test blueprint, response probability, item pool refinement, and passage-related concerns. Each of these is discussed in detail on the following pages.

Ability Estimates

As described in the previous section, DRC BEACON item pools are calibrated using item response theory models and are vertically linked across grades. The CAT algorithm has access to each item's operational parameters in the item pool. After each item's response, ability estimates and associated standard errors are calculated via maximum likelihood estimation for the total test and for each reporting category. In the case of zero (all items incorrect) and perfect (all items correct) scores, a correction factor is applied before computing the relevant maximum likelihood estimates. A fractional value is added to a zero score and subtracted.

Item Selection

Note that rather than identifying a single item that best matches the student's performance, one item is randomly selected from among a set of items that are all near the student's current ability estimate and

meet other configuration requirements discussed in this section. Randomly selecting items from a set of items that meet the content and psychometric specifications of the algorithm helps minimize repeated exposure to content throughout the test administrations.

Test Blueprint

The CAT algorithm used within DRC BEACON closely resembles a modified constrained CAT (MCCAT) design (Leung et al., 2003). The CAT algorithm is configured with upper and lower bounds that specify the minimum and maximum numbers of items that will be administered to students for both the total test and the reporting categories. The algorithm keeps track of what parts of the blueprint have been filled and what parts remain relative to the blueprint configurations as it proceeds through the assessment.

Response Probability

The CAT algorithm used within DRC BEACON targets items where the student has response probability (RP) of answering correctly, based on the ability estimate and item parameters associated with the item using the item response theory model discussed in the previous section. Theoretically, the RP of 0.5 is the most efficient value to use within adaptive testing because item information is maximized at this probability. That is, selecting items that the student has a 50% chance of answering correctly will produce the smallest standard error for any given number of items.

Item Pool Refinement

The CAT algorithm includes configurable elements that allow for the refinement of the pool used in item selection. Two configurable elements are listed here.

> **Restrict pool**—The ability to restrict the available item pool by grade or course at various points in the test. For example, off grade/level items are not available in the first segment of an assessment. This restriction can then be removed in subsequent segments to allow content aligned with adjacent grade levels to be included.
>
> **Favor items**—The ability to favor items that are close to the student's grade when evaluating items near a student's estimated score. For example, if a student is in grade 8 and the item selection routinely finds appropriate items (in terms of difficulty) in grades 4, 5, 6, 7, and 8, the CAT algorithm can favor items at or close to grade 8. It is possible that no items near a student's grade are appropriate in terms of difficulty. In that case, the CAT algorithm will select items further away from the student's grade but appropriate based on item difficulty.

The difference between restricting the pool and favoring items is that when the pool is restricted, some items cannot be selected. With favoring, all non-restricted items are eligible for administration, but they are made more or less likely to be selected based on closeness to student grade. DRC BEACON uses both restriction and favoring rules throughout the test administration.

Passage Considerations

The DRC BEACON tests in ELA include many reading and listening items that are passage based. These passages have between 3 and 12 associated items. The CAT algorithm does not require that all items associated with a passage be administered. Instead, it evaluates all possible combinations of items associated with a passage. Item sequencing within a passage is preserved when items are presented to the student. For example, if a six-item passage is selected and items 1 and 4 are not administered, then items 2, 3, 5, and 6 will be administered with the passage in sequence.

The configurable elements of passage-based CAT tests include the following:

Passage minimum percentage—Define the minimum percentage of the items associated with a passage that need to be used. For example, if the passage minimum percent is set at 80, then the selection routine will consider combinations such as 4 of 5 (80%), 5 of 6 (83%), and 6 of 6 (100%). It will not consider combinations such as 1 of 2 (50%), 3 of 4 (75%), 3 of 5 (60%), etc. Near the end of a test, the passage minimum percent constraint may need to be loosened to meet content constraints such as number of items per reporting category.

Passage evaluation criteria—Multiple factors are considered when evaluating and ranking each passage combination to determine the best combination to administer. They include the following:

- Percentage of items associated with the passage used; the higher the percent, the higher the combination is ranked
- Number of items associated with the passage used; the higher the number, the higher the combination is ranked
- Distance between items' difficulties and student's estimated score; the smaller the distance, the higher the combination is ranked
- Distance between the items' grade levels and the student's grade level; the smaller the distance, the higher the combination is ranked

Different weights may be assigned to each of the factors. For example, if all the weight is put on the number of items used, then the algorithm will select the passages with the closest number of items and administer all of them until the maximum number of items is reached.

Test Navigation

Many versions of computer-adaptive tests do not allow students to skip items in the test or back up to previously answered items and change answers due to some complicating factors.

If students were allowed to skip items, the CAT algorithm would need to select additional items without any additional information (no change to ability estimates). Taken to the extreme, a student with no prior scores who skipped every item starting with the first would receive an entire test of items with average difficulty. It would not be adaptive at all.

If students were allowed to back up and change answers, ability estimates would need to be recalculated when answers were changed. This additional information could be used to select additional items but would not change previously selected items. Also, if students were allowed to back up in the test, additional considerations would need to be put in place to ensure that the answer to one item does not cue another.

Generally speaking, DRC BEACON tests do not allow skipping items or backing up and changing answers. However, in passages that are selected to measure reading and listening, students can skip and go back to items. For example, when presented with a passage and five associated items, the student does not have to answer questions one through five in order without skipping. However, if a student tries to navigate to the next passage without answering all five items associated with the first passage, the test engine will prompt the student to answer all items and will not move on to the next passage until all are answered.

Termination

CAT algorithms can be configured for fixed length or variable length testing. With fixed length testing, the test ends when a student has taken a predefined number of items total and in each reporting category. With variable length testing, the algorithm stops administering items from a reporting category when one of two conditions is satisfied—when a student has taken more than a predefined minimum number of items in that reporting category and the standard error is below a predefined threshold or when a student has taken a predefined maximum number of items in that reporting category. The test ends when one of the two conditions above is satisfied for each of the reporting categories. Note that with both fixed length and variable length tests, there is no requirement that the predefined number of items in reporting categories be equal. Fixed length testing is currently specified for use within DRC BEACON.

Embedded Field Test Items

Over time, additional items will be needed to replenish the DRC BEACON item pools. Embedding field test items within an operational administration is advantageous for two reasons. First, sufficient item response data can be gathered without the time and expense of a separate stand-alone administration. Second, it allows the new items to be placed on the existing operational scale.

BEACON regularly includes embedded field test items within each test administration. For each embedded field test event, the factors considered when determining the number of field test items to embed included the number of items to be field-tested, the expected number of students testing, and

the desired number of students per item for field test analyses. In mathematics, field test items were randomly assigned to fixed positions spread throughout the operational test. In ELA, a field test passage was randomly assigned near the middle of the test and students took all the items associated with the passage. In all content areas, the positions of field test items were unknown to students. Field test items were not clustered at the end of the test to avoid any fatigue effect when placing the items on the operational scale.

BEACON CAT Configuration

Elements of the CAT configurations that support the DRC BEACON administration are reported below for the comprehensive ELA and mathematics assessments. Details regarding the CAT configurations that support the administration of DRC BEACON testlets are also provided. Note that an extensive set of simulations were run with each configuration and the results are provided in the subsequent section.

CAT Configuration—Full ELA Assessments

The ELA assessment is configured with respect to seven reporting categories. Each student will take between 8 and 10 operational items per reporting category resulting in a total test of between 56 and 61 operational items. With no prior information about a student, the starting point will be an item of average difficulty based on grade level. For example, a grade 7 student will start with an item that is near the average difficulty of grade 7 items. Items are selected that have a response probability of 0.5, meaning a student has a 50% chance of answering correctly. The CAT algorithm will stop administering items when a student has taken 8 to 10 operational items in all seven reporting categories.

Functionality is used to restrict the pool and to favor items close to a student's grade. The pool restrictions are listed below.

- No passages are administered in the initial locator segment.
- Literary and informational passages and associated items are administered in approximately equal quantity.
- Passage minimum percentage is set at 66%. That is, whenever possible, only passage combinations that use 66% or more of the associated items are used. (Near the end of a test, the passage minimum percent constraint may need to be loosened to meet content constraints.) Many simulations were run to arrive at this percentage. On one hand, testing time and reading load should be minimized. Therefore, students should not have to read long passages for only one or two items. On the other hand, using all items associated with a passage may not be desirable since some items are far from a student's estimated score. Given a limited number of items, those that are either too easy or too hard should not be used.
- No off-grade items will be administered in the first 6 items.
- Off-grade items from adjacent grades are allowed in items 7 through 30.
- Off-grade items from two adjacent grades are allowed in items 31 and beyond.

A number of testlets are also available for the DRC BEACON ELA assessments. Testlets allow students to take a test that focuses on specific aspects of content. The following six testlets are available in ELA:

Reading/Writing Only, Reading Only, Listening Only, Writing—Text Types and Purposes, Writing—Conventions, and Writing—Research.

The configuration for the Reading/Writing Only testlet maintains the same set of rules as the full adaptive assessment, except the pool is restricted to exclude Listening items from consideration. Similarly, the configuration supporting the Reading Only testlet delivers an adaptive assessment that includes only Reading content and excludes all Writing and Listening content. The configurations supporting the Listening Only testlet and the three Writing testlets are extended from 8 items each to 10 items each to provide a more accurate ability estimate than what is possible within the full ELA assessments.

CAT Configuration—Mathematics Assessments

The DRC BEACON mathematics assessment is configured with respect to four reporting categories. Each student will take 8 operational items per reporting category resulting in a total test of 32 operational items. With no prior information about a student, the starting point will be an item of average difficulty based on grade level. For example, a grade 7 student will start with an item near the average difficulty of grade 7 items. Items are selected that have a response probability of 0.5, meaning a student has a 50% chance of answering correctly. The CAT algorithm will stop administering items when the student has taken a total of 32 items.

Functionality is used to restrict the pool and to favor items close to a student's grade. Additional pool restrictions are listed below.

- No off-grade items will be administered in the first 5 items.
- Off-grade items from adjacent grades are allowed in items 6 through 22.
- Off-grade items from two adjacent grades are allowed in items 23 and beyond.
- Four field test items are administered within the assessment. The field test items are restricted based on grade level.

Mathematics testlets are also available within DRC BEACON. Testlets allow students to take a test that focuses on one of the four reporting categories: Algebra, Numbers and Quantity, Measurement and Data, and Geometry. Given that the content is limited to a single reporting category when testlets are administered, the number of items for each reporting category is increased from 8 items to 10 items. This allows for more precise estimates of ability within a reporting category than are available within the full test. Similar functionality is used to restrict the pool to the reporting category and to favor items close to a student grade when testlets are administered.

Simulation Results

Once all content and psychometric specifications were configured for DRC BEACON, extensive simulated test administrations were conducted prior to the first live test administration. The simulation was designed to mimic an actual test administration using 3,000 computer generated students with known ability levels. The simulation tested the functionality of the adaptive testing system and item pool across the full range of student proficiency in the following areas: test blueprint coverage, item exposure, student ability estimation, and the standard errors of measurement. The simulation results are provided for the full DRC BEACON assessments and the testlets in tables in the following pages. Overall, the results are as expected and meet the acceptable psychometric requirements given the available item pool.

Test Blueprint Coverage

An essential requirement of adaptive testing is that each test administration must meet the test blueprint and associated test specifications to assure the comparability of student scores. In addition to verifying that student administrations are aligned to the test blueprint, DRC BEACON uses the simulation tool to administration ordered patterns items and/or item sets to assure the appropriateness of each test event.

The simulation results show that every student received the correct number of items as configured. Note that for ELA test administrations that include Reading or Listening, the total number of items administered will vary as a function of how many items are administered for specific passages. The results indicate that when administering DRC BEACON, the CAT engine offered students the expected number of items and points, the expected number of passages, and a reasonable number of items per passage.

Item Exposure

A common concern when implementing adaptive tests is the exposure rate of the items. It is important to control the item exposure rate while balancing the other constraints of the CAT. Table 28 shows the item exposure rates for ELA and mathematics. The table provides the number and percentage of items for each of five exposure rate categories. For example, an exposure rate of [0.0, 0.1] means that 0 to 10% of the students took that item. For both ELA and mathematics, most of the items have low exposure rates and are categorized with an exposure rate between 0.0 and 0.1. For all grades and content areas, no more than 3% of items have an exposure rate greater than 0.4.

Table 28. Summary of Item Exposure Rate ELA and Mathematics Full Tests

		ELA		Mathematics	
Level	Item Exposure	Number of Items	Proportion of Items	Number of Items	Proportion of Items
3	[0.0, 0.1]	552	0.74	373	0.74
	(0.1, 0.2]	99	0.13	66	0.13
	(0.2, 0.3]	43	0.06	45	0.09
	(0.3, 0.4]	32	0.04	17	0.03
	> 0.4	24	0.03	2	0.00
4	[0.0, 0.1]	803	0.79	544	0.81
	(0.1, 0.2]	141	0.14	89	0.13
	(0.2, 0.3]	47	0.05	30	0.04
	(0.3, 0.4]	13	0.01	5	0.01
	> 0.4	9	0.01	0	0.00
5	[0.0, 0.1]	1,060	0.84	683	0.83
	(0.1, 0.2]	132	0.10	112	0.14
	(0.2, 0.3]	52	0.04	22	0.03
	(0.3, 0.4]	18	0.01	1	0.00
	> 0.4	4	0.00	0	0.00
6	[0.0, 0.1]	1,091	0.86	677	0.82
	(0.1, 0.2]	115	0.09	110	0.13
	(0.2, 0.3]	40	0.03	28	0.03
	(0.3, 0.4]	15	0.01	6	0.01
	> 0.4	8	0.01	0	0.00
7	[0.0, 0.1]	817	0.81	519	0.80
	(0.1, 0.2]	110	0.11	97	0.15
	(0.2, 0.3]	55	0.05	30	0.05
	(0.3, 0.4]	22	0.02	3	0.00
	> 0.4	10	0.01	0	0.00
8	[0.0, 0.1]	550	0.72	343	0.72
	(0.1, 0.2]	135	0.18	85	0.18
	(0.2, 0.3]	33	0.04	44	0.09
	(0.3, 0.4]	19	0.02	2	0.00
	> 0.4	25	0.03	0	0.00

Table 29. Summary of ELA Full Tests Reporting Category Item Exposure Rate

		Reading: Key Ideas and Details		Reading: Craft Structure/Integration of Knowledge and Ideas		Reading: Vocabulary Acquisition and Use	
Level	Item Exposure	Number of Items	Proportion of Items	Number of Items	Proportion of Items	Number of Items	Proportion of Items
3	[0.0, 0.1]	363	0.93	228	0.87	126	0.83
	(0.1, 0.2]	15	0.04	18	0.07	6	0.04
	(0.2, 0.3]	5	0.01	8	0.03	9	0.06
	(0.3, 0.4]	5	0.01	3	0.01	7	0.05
	> 0.4	4	0.01	4	0.02	3	0.02
4	[0.0, 0.1]	354	0.90	227	0.87	122	0.81
	(0.1, 0.2]	30	0.08	24	0.09	13	0.09
	(0.2, 0.3]	8	0.02	7	0.03	11	0.07
	(0.3, 0.4]	0	0.00	2	0.01	4	0.03
	> 0.4	0	0.00	1	0.00	1	0.01
5	[0.0, 0.1]	359	0.92	223	0.85	122	0.81
	(0.1, 0.2]	27	0.07	25	0.10	17	0.11
	(0.2, 0.3]	4	0.01	12	0.05	3	0.02
	(0.3, 0.4]	1	0.00	1	0.00	8	0.05
	> 0.4	1	0.00	0	0.00	1	0.01
6	[0.0, 0.1]	359	0.92	233	0.89	130	0.86
	(0.1, 0.2]	25	0.06	23	0.09	8	0.05
	(0.2, 0.3]	6	0.02	3	0.01	4	0.03
	(0.3, 0.4]	2	0.01	0	0.00	7	0.05
	> 0.4	0	0.00	2	0.01	2	0.01
7	[0.0, 0.1]	362	0.92	230	0.88	128	0.85
	(0.1, 0.2]	17	0.04	16	0.06	8	0.05
	(0.2, 0.3]	9	0.02	8	0.03	7	0.05
	(0.3, 0.4]	3	0.01	5	0.02	5	0.03
	> 0.4	1	0.00	2	0.01	3	0.02
8	[0.0, 0.1]	353	0.90	225	0.86	128	0.85
	(0.1, 0.2]	30	0.08	26	0.10	8	0.05
	(0.2, 0.3]	7	0.02	5	0.02	3	0.02
	(0.3, 0.4]	2	0.01	2	0.01	7	0.05
	> 0.4	0	0.00	3	0.01	5	0.03

Table 29. Summary of ELA Full Tests Reporting Category Item Exposure Rate (continued)

		Writing – Text Types		Writing – Conventions		Writing – Research		Listening	
Level	Item Exposure	Number of Items	Proportion of Items	Number of Items	Proportion of Items	Number of Items	Proportion of Items	Number of Items	Proportion of Items
3	[0.0, 0.1]	170	0.85	126	0.85	89	0.78	212	0.86
	(0.1, 0.2]	15	0.08	7	0.05	9	0.08	29	0.12
	(0.2, 0.3]	6	0.03	6	0.04	6	0.05	3	0.01
	(0.3, 0.4]	4	0.02	6	0.04	5	0.04	2	0.01
	> 0.4	4	0.02	4	0.03	5	0.04	0	0.00
4	[0.0, 0.1]	172	0.86	127	0.85	83	0.73	217	0.88
	(0.1, 0.2]	17	0.09	11	0.07	17	0.15	29	0.12
	(0.2, 0.3]	7	0.04	6	0.04	8	0.07	0	0.00
	(0.3, 0.4]	2	0.01	2	0.01	3	0.03	0	0.00
	> 0.4	1	0.01	3	0.02	3	0.03	0	0.00
5	[0.0, 0.1]	177	0.89	117	0.79	83	0.73	225	0.91
	(0.1, 0.2]	13	0.07	24	0.16	13	0.11	13	0.05
	(0.2, 0.3]	7	0.04	5	0.03	13	0.11	8	0.03
	(0.3, 0.4]	2	0.01	3	0.02	3	0.03	0	0.00
	> 0.4	0	0.00	0	0.00	2	0.02	0	0.00
6	[0.0, 0.1]	178	0.89	122	0.82	86	0.75	226	0.92
	(0.1, 0.2]	13	0.07	16	0.11	17	0.15	13	0.05
	(0.2, 0.3]	8	0.04	9	0.06	4	0.04	6	0.02
	(0.3, 0.4]	0	0.00	1	0.01	4	0.04	1	0.00
	> 0.4	0	0.00	1	0.01	3	0.03	0	0.00
7	[0.0, 0.1]	173	0.87	119	0.80	85	0.75	218	0.89
	(0.1, 0.2]	17	0.09	17	0.11	15	0.13	20	0.08
	(0.2, 0.3]	6	0.03	10	0.07	7	0.06	8	0.03
	(0.3, 0.4]	2	0.01	1	0.01	6	0.05	0	0.00
	> 0.4	1	0.01	2	0.01	1	0.01	0	0.00
8	[0.0, 0.1]	170	0.85	123	0.83	88	0.77	213	0.87
	(0.1, 0.2]	18	0.09	14	0.09	13	0.11	26	0.11
	(0.2, 0.3]	7	0.04	4	0.03	3	0.03	4	0.02
	(0.3, 0.4]	2	0.01	2	0.01	2	0.02	2	0.01
	> 0.4	2	0.01	6	0.04	8	0.07	1	0.00

Table 30. Summary of Mathematics Full Tests Reporting Category Item Exposure Rate

		Algebra		Number & Quantity		Measurement & Data		Geometry	
Level	Item Exposure	Number of Items	Proportion of Items	Number of Items	Proportion of Items	Number of Items	Proportion of Items	Number of Items	Proportion of Items
3	[0.0, 0.1]	233	0.87	305	0.89	177	0.86	132	0.83
	(0.1, 0.2]	22	0.08	27	0.08	10	0.05	7	0.04
	(0.2, 0.3]	14	0.05	11	0.03	10	0.05	10	0.06
	(0.3, 0.4]	0	0.00	0	0.00	9	0.04	8	0.05
	> 0.4	0	0.00	0	0.00	0	0.00	2	0.01
4	[0.0, 0.1]	243	0.90	309	0.90	175	0.85	126	0.79
	(0.1, 0.2]	16	0.06	32	0.09	22	0.11	19	0.12
	(0.2, 0.3]	10	0.04	2	0.01	8	0.04	10	0.06
	(0.3, 0.4]	0	0.00	0	0.00	1	0.00	4	0.03
	> 0.4	0	0.00	0	0.00	0	0.00	0	0.00
5	[0.0, 0.1]	235	0.87	316	0.92	169	0.82	122	0.77
	(0.1, 0.2]	28	0.10	27	0.08	28	0.14	29	0.18
	(0.2, 0.3]	6	0.02	0	0.00	9	0.04	7	0.04
	(0.3, 0.4]	0	0.00	0	0.00	0	0.00	1	0.01
	> 0.4	0	0.00	0	0.00	0	0.00	0	0.00
6	[0.0, 0.1]	232	0.86	306	0.89	169	0.82	126	0.79
	(0.1, 0.2]	28	0.10	33	0.10	30	0.15	19	0.12
	(0.2, 0.3]	9	0.03	4	0.01	5	0.02	10	0.06
	(0.3, 0.4]	0	0.00	0	0.00	2	0.01	4	0.03
	> 0.4	0	0.00	0	0.00	0	0.00	0	0.00
7	[0.0, 0.1]	238	0.88	311	0.91	174	0.84	124	0.78
	(0.1, 0.2]	23	0.09	28	0.08	21	0.10	25	0.16
	(0.2, 0.3]	8	0.03	4	0.01	11	0.05	7	0.04
	(0.3, 0.4]	0	0.00	0	0.00	0	0.00	3	0.02
	> 0.4	0	0.00	0	0.00	0	0.00	0	0.00
8	[0.0, 0.1]	235	0.87	312	0.91	172	0.83	127	0.80
	(0.1, 0.2]	28	0.10	22	0.06	22	0.11	13	0.08
	(0.2, 0.3]	6	0.02	9	0.03	10	0.05	19	0.12
	(0.3, 0.4]	0	0.00	0	0.00	2	0.01	0	0.00
	> 0.4	0	0.00	0	0.00	0	0.00	0	0.00

Table 31. Summary of ELA Reporting Category Reading and Writing Testlet Item Exposure Rate

		Total		Reading: Key Ideas and Details		Reading: Craft Structure/Integration of Knowledge and Ideas		Reading: Vocabulary Acquisition and Use	
Level	Item Exposure	Number of Items	Proportion of Items	Number of Items	Proportion of Items	Number of Items	Proportion of Items	Number of Items	Proportion of Items
3	[0.0, 0.1]	1,113	0.88	364	0.93	233	0.89	127	0.84
	(0.1, 0.2]	63	0.05	15	0.04	12	0.05	9	0.06
	(0.2, 0.3]	34	0.03	5	0.01	9	0.03	6	0.04
	(0.3, 0.4]	25	0.02	2	0.01	3	0.01	3	0.02
	> 0.4	31	0.02	6	0.02	4	0.02	6	0.04
4	[0.0, 0.1]	1,091	0.86	356	0.91	230	0.88	126	0.83
	(0.1, 0.2]	89	0.07	21	0.05	17	0.07	7	0.05
	(0.2, 0.3]	45	0.04	12	0.03	7	0.03	6	0.04
	(0.3, 0.4]	25	0.02	3	0.01	5	0.02	9	0.06
	> 0.4	16	0.01	0	0.00	2	0.01	3	0.02
5	[0.0, 0.1]	1,083	0.86	360	0.92	228	0.87	124	0.82
	(0.1, 0.2]	105	0.08	16	0.04	18	0.07	15	0.10
	(0.2, 0.3]	41	0.03	10	0.03	9	0.03	4	0.03
	(0.3, 0.4]	26	0.02	4	0.01	4	0.02	4	0.03
	> 0.4	11	0.01	2	0.01	2	0.01	4	0.03
6	[0.0, 0.1]	1,090	0.86	358	0.91	231	0.89	132	0.87
	(0.1, 0.2]	109	0.09	22	0.06	19	0.07	5	0.03
	(0.2, 0.3]	36	0.03	9	0.02	7	0.03	6	0.04
	(0.3, 0.4]	17	0.01	3	0.01	3	0.01	1	0.01
	> 0.4	14	0.01	0	0.00	1	0.00	7	0.05
7	[0.0, 0.1]	1,099	0.87	361	0.92	228	0.87	132	0.87
	(0.1, 0.2]	81	0.06	18	0.05	13	0.05	5	0.03
	(0.2, 0.3]	42	0.03	4	0.01	11	0.04	5	0.03
	(0.3, 0.4]	27	0.02	6	0.02	7	0.03	3	0.02
	> 0.4	17	0.01	3	0.01	2	0.01	6	0.04
8	[0.0, 0.1]	1,096	0.87	355	0.91	225	0.86	129	0.85
	(0.1, 0.2]	81	0.06	22	0.06	17	0.07	8	0.05
	(0.2, 0.3]	40	0.03	10	0.03	12	0.05	3	0.02
	(0.3, 0.4]	19	0.02	3	0.01	5	0.02	3	0.02
	> 0.4	30	0.02	2	0.01	2	0.01	8	0.05

Table 31. Summary of ELA Reporting Category Reading and Writing Testlet Item Exposure Rate (continued)

		Writing – Text Types and Purposes		Writing – Conventions of Standard English		Writing – Research	
Level	Item Exposure	Number of Items	Proportion of Items	Number of Items	Proportion of Items	Number of Items	Proportion of Items
3	[0.0, 0.1]	174	0.87	126	0.85	89	0.78
	(0.1, 0.2]	11	0.06	6	0.04	10	0.09
	(0.2, 0.3]	5	0.03	7	0.05	2	0.02
	(0.3, 0.4]	4	0.02	6	0.04	7	0.06
	> 0.4	5	0.03	4	0.03	6	0.05
4	[0.0, 0.1]	171	0.86	126	0.85	82	0.72
	(0.1, 0.2]	16	0.08	11	0.07	17	0.15
	(0.2, 0.3]	6	0.03	4	0.03	10	0.09
	(0.3, 0.4]	5	0.03	2	0.01	1	0.01
	> 0.4	1	0.01	6	0.04	4	0.04
5	[0.0, 0.1]	172	0.86	115	0.77	84	0.74
	(0.1, 0.2]	20	0.10	24	0.16	12	0.11
	(0.2, 0.3]	3	0.02	5	0.03	10	0.09
	(0.3, 0.4]	4	0.02	4	0.03	6	0.05
	> 0.4	0	0.00	1	0.01	2	0.02
6	[0.0, 0.1]	168	0.84	118	0.79	83	0.73
	(0.1, 0.2]	23	0.12	22	0.15	18	0.16
	(0.2, 0.3]	5	0.03	4	0.03	5	0.04
	(0.3, 0.4]	3	0.02	4	0.03	3	0.03
	> 0.4	0	0.00	1	0.01	5	0.04
7	[0.0, 0.1]	173	0.87	121	0.81	84	0.74
	(0.1, 0.2]	14	0.07	17	0.11	14	0.12
	(0.2, 0.3]	9	0.05	4	0.03	9	0.08
	(0.3, 0.4]	2	0.01	5	0.03	4	0.04
	> 0.4	1	0.01	2	0.01	3	0.03
8	[0.0, 0.1]	172	0.86	125	0.84	90	0.79
	(0.1, 0.2]	16	0.08	10	0.07	8	0.07
	(0.2, 0.3]	4	0.02	4	0.03	7	0.06
	(0.3, 0.4]	3	0.02	4	0.03	1	0.01
	> 0.4	4	0.02	6	0.04	8	0.07

Table 32. Summary of ELA Reporting Category Reading Testlets Item Exposure Rate

		Total		Reading: Key Ideas and Details		Reading: Craft Structure/Integration of Knowledge and Ideas		Reading: Vocabulary Acquisition and Use	
Level	Item Exposure	Number of Items	Proportion of Items	Number of Items	Proportion of Items	Number of Items	Proportion of Items	Number of Items	Proportion of Items
3	[0.0, 0.1]	710	0.88	359	0.92	224	0.86	127	0.84
	(0.1, 0.2]	58	0.07	20	0.05	27	0.10	11	0.07
	(0.2, 0.3]	15	0.02	8	0.02	4	0.02	3	0.02
	(0.3, 0.4]	15	0.02	3	0.01	5	0.02	7	0.05
	> 0.4	6	0.01	2	0.01	1	0.00	3	0.02
4	[0.0, 0.1]	717	0.89	361	0.92	231	0.89	125	0.83
	(0.1, 0.2]	60	0.07	25	0.06	23	0.09	12	0.08
	(0.2, 0.3]	11	0.01	3	0.01	3	0.01	5	0.03
	(0.3, 0.4]	13	0.02	2	0.01	3	0.01	8	0.05
	> 0.4	3	0.00	1	0.00	1	0.00	1	0.01
5	[0.0, 0.1]	720	0.90	364	0.93	227	0.87	129	0.85
	(0.1, 0.2]	48	0.06	14	0.04	23	0.09	11	0.07
	(0.2, 0.3]	21	0.03	11	0.03	7	0.03	3	0.02
	(0.3, 0.4]	9	0.01	3	0.01	4	0.02	2	0.01
	> 0.4	6	0.01	0	0.00	0	0.00	6	0.04
6	[0.0, 0.1]	729	0.91	367	0.94	230	0.88	132	0.87
	(0.1, 0.2]	44	0.05	17	0.04	21	0.08	6	0.04
	(0.2, 0.3]	16	0.02	6	0.02	7	0.03	3	0.02
	(0.3, 0.4]	6	0.01	0	0.00	1	0.00	5	0.03
	> 0.4	9	0.01	2	0.01	2	0.01	5	0.03
7	[0.0, 0.1]	730	0.91	364	0.93	234	0.90	132	0.87
	(0.1, 0.2]	44	0.05	19	0.05	17	0.07	8	0.05
	(0.2, 0.3]	13	0.02	4	0.01	8	0.03	1	0.01
	(0.3, 0.4]	9	0.01	3	0.01	0	0.00	6	0.04
	> 0.4	8	0.01	2	0.01	2	0.01	4	0.03
8	[0.0, 0.1]	720	0.90	356	0.91	232	0.89	132	0.87
	(0.1, 0.2]	58	0.07	31	0.08	22	0.08	5	0.03
	(0.2, 0.3]	8	0.01	1	0.00	4	0.02	3	0.02
	(0.3, 0.4]	5	0.01	3	0.01	1	0.00	1	0.01
	> 0.4	13	0.02	1	0.00	2	0.01	10	0.07

Table 33. Summary of ELA Reporting Category Writing – Text Types & Purposes Testlet Item Exposure Rate

		Writing – Text Types and Purposes	
Level	**Item Exposure**	**Number of Items**	**Proportion of Items**
3	[0.0, 0.1]	168	0.84
	(0.1, 0.2]	15	0.08
	(0.2, 0.3]	7	0.04
	(0.3, 0.4]	1	0.01
	> 0.4	8	0.04
4	[0.0, 0.1]	164	0.82
	(0.1, 0.2]	24	0.12
	(0.2, 0.3]	2	0.01
	(0.3, 0.4]	2	0.01
	> 0.4	7	0.04
5	[0.0, 0.1]	169	0.85
	(0.1, 0.2]	21	0.11
	(0.2, 0.3]	1	0.01
	(0.3, 0.4]	0	0.00
	> 0.4	8	0.04
6	[0.0, 0.1]	168	0.84
	(0.1, 0.2]	21	0.11
	(0.2, 0.3]	1	0.01
	(0.3, 0.4]	1	0.01
	> 0.4	8	0.04
7	[0.0, 0.1]	171	0.86
	(0.1, 0.2]	16	0.08
	(0.2, 0.3]	4	0.02
	(0.3, 0.4]	1	0.01
	> 0.4	7	0.04
8	[0.0, 0.1]	162	0.81
	(0.1, 0.2]	25	0.13
	(0.2, 0.3]	4	0.02
	(0.3, 0.4]	1	0.01
	> 0.4	7	0.04

Table 34. Summary of ELA Reporting Category Writing – Conventions of Standard English Testlet Item Exposure Rate

		Writing – Conventions of Standard English	
Level	**Item Exposure**	**Number of Items**	**Proportion of Items**
3	[0.0, 0.1]	118	0.79
	(0.1, 0.2]	14	0.09
	(0.2, 0.3]	8	0.05
	(0.3, 0.4]	0	0.00
	> 0.4	9	0.06
4	[0.0, 0.1]	121	0.81
	(0.1, 0.2]	18	0.12
	(0.2, 0.3]	2	0.01
	(0.3, 0.4]	0	0.00
	> 0.4	8	0.05
5	[0.0, 0.1]	117	0.79
	(0.1, 0.2]	21	0.14
	(0.2, 0.3]	2	0.01
	(0.3, 0.4]	2	0.01
	> 0.4	7	0.05
6	[0.0, 0.1]	111	0.74
	(0.1, 0.2]	27	0.18
	(0.2, 0.3]	3	0.02
	(0.3, 0.4]	1	0.01
	> 0.4	7	0.05
7	[0.0, 0.1]	117	0.79
	(0.1, 0.2]	19	0.13
	(0.2, 0.3]	5	0.03
	(0.3, 0.4]	0	0.00
	> 0.4	8	0.05
8	[0.0, 0.1]	118	0.79
	(0.1, 0.2]	16	0.11
	(0.2, 0.3]	6	0.04
	(0.3, 0.4]	1	0.01
	> 0.4	8	0.05

Table 35. Summary of ELA Reporting Category Writing – Research Testlet Item Exposure Rate

		Writing – Research	
Level	**Item Exposure**	**Number of Items**	**Proportion of Items**
3	[0.0, 0.1]	89	0.78
	(0.1, 0.2]	8	0.07
	(0.2, 0.3]	3	0.03
	(0.3, 0.4]	4	0.04
	> 0.4	10	0.09
4	[0.0, 0.1]	81	0.71
	(0.1, 0.2]	15	0.13
	(0.2, 0.3]	9	0.08
	(0.3, 0.4]	1	0.01
	> 0.4	8	0.07
5	[0.0, 0.1]	84	0.74
	(0.1, 0.2]	14	0.12
	(0.2, 0.3]	6	0.05
	(0.3, 0.4]	2	0.02
	> 0.4	8	0.07
6	[0.0, 0.1]	86	0.75
	(0.1, 0.2]	15	0.13
	(0.2, 0.3]	4	0.04
	(0.3, 0.4]	2	0.02
	> 0.4	7	0.06
7	[0.0, 0.1]	82	0.72
	(0.1, 0.2]	19	0.17
	(0.2, 0.3]	4	0.04
	(0.3, 0.4]	1	0.01
	> 0.4	8	0.07
8	[0.0, 0.1]	88	0.77
	(0.1, 0.2]	9	0.08
	(0.2, 0.3]	5	0.04
	(0.3, 0.4]	3	0.03
	> 0.4	9	0.08

Table 36. Summary of ELA Reporting Category Listening Testlet Item Exposure Rate

		Listening	
Level	**Item Exposure**	**Number of Items**	**Proportion of Items**
3	[0.0, 0.1]	209	0.85
	(0.1, 0.2]	22	0.09
	(0.2, 0.3]	11	0.04
	(0.3, 0.4]	1	0.00
	> 0.4	3	0.01
4	[0.0, 0.1]	205	0.83
	(0.1, 0.2]	32	0.13
	(0.2, 0.3]	7	0.03
	(0.3, 0.4]	2	0.01
	> 0.4	0	0.00
5	[0.0, 0.1]	203	0.83
	(0.1, 0.2]	30	0.12
	(0.2, 0.3]	13	0.05
	(0.3, 0.4]	0	0.00
	> 0.4	0	0.00
6	[0.0, 0.1]	212	0.86
	(0.1, 0.2]	21	0.09
	(0.2, 0.3]	8	0.03
	(0.3, 0.4]	5	0.02
	> 0.4	0	0.00
7	[0.0, 0.1]	214	0.87
	(0.1, 0.2]	18	0.07
	(0.2, 0.3]	7	0.03
	(0.3, 0.4]	7	0.03
	> 0.4	0	0.00
8	[0.0, 0.1]	208	0.85
	(0.1, 0.2]	26	0.11
	(0.2, 0.3]	6	0.02
	(0.3, 0.4]	1	0.00
	> 0.4	5	0.02

Table 37. Summary of Mathematics Reporting Category Testlets Item Exposure Rate

		Algebra		Number & Quantity		Measurement & Data		Geometry	
Level	Item Exposure	Number of Items	Proportion of Items	Number of Items	Proportion of Items	Number of Items	Proportion of Items	Number of Items	Proportion of Items
3	[0.0, 0.1]	463	0.92	463	0.92	473	0.94	475	0.94
	(0.1, 0.2]	15	0.03	18	0.04	7	0.01	3	0.01
	(0.2, 0.3]	23	0.05	22	0.04	12	0.02	10	0.02
	(0.3, 0.4]	2	0.00	0	0.00	6	0.01	8	0.02
	> 0.4	0	0.00	0	0.00	5	0.01	7	0.01
4	[0.0, 0.1]	233	0.87	309	0.90	171	0.83	119	0.75
	(0.1, 0.2]	22	0.08	19	0.06	23	0.11	18	0.11
	(0.2, 0.3]	14	0.05	15	0.04	12	0.06	19	0.12
	(0.3, 0.4]	0	0.00	0	0.00	0	0.00	2	0.01
	> 0.4	0	0.00	0	0.00	0	0.00	1	0.01
5	[0.0, 0.1]	227	0.84	306	0.89	162	0.79	120	0.75
	(0.1, 0.2]	31	0.12	29	0.08	29	0.14	20	0.13
	(0.2, 0.3]	11	0.04	8	0.02	15	0.07	19	0.12
	(0.3, 0.4]	0	0.00	0	0.00	0	0.00	0	0.00
	> 0.4	0	0.00	0	0.00	0	0.00	0	0.00
6	[0.0, 0.1]	227	0.84	304	0.89	165	0.80	120	0.75
	(0.1, 0.2]	28	0.10	25	0.07	29	0.14	17	0.11
	(0.2, 0.3]	14	0.05	14	0.04	12	0.06	20	0.13
	(0.3, 0.4]	0	0.00	0	0.00	0	0.00	2	0.01
	> 0.4	0	0.00	0	0.00	0	0.00	0	0.00
7	[0.0, 0.1]	235	0.87	308	0.90	164	0.80	115	0.72
	(0.1, 0.2]	25	0.09	24	0.07	26	0.13	26	0.16
	(0.2, 0.3]	9	0.03	11	0.03	16	0.08	17	0.11
	(0.3, 0.4]	0	0.00	0	0.00	0	0.00	1	0.01
	> 0.4	0	0.00	0	0.00	0	0.00	0	0.00
8	[0.0, 0.1]	231	0.86	310	0.90	163	0.79	124	0.78
	(0.1, 0.2]	22	0.08	17	0.05	25	0.12	11	0.07
	(0.2, 0.3]	16	0.06	16	0.05	15	0.07	19	0.12
	(0.3, 0.4]	0	0.00	0	0.00	3	0.01	5	0.03
	> 0.4	0	0.00	0	0.00	0	0.00	0	0.00

Evaluating Student Ability Estimation

DRC conducted simulation studies for DRC BEACON using ability estimates sampled from actual DRC BEACON administrations. To estimate the examinee ability in the simulation study, maximum likelihood estimation (MLE) was utilized. To limit extreme values in the score range, the test scoring algorithm used the highest and lowest obtainable scale scores (i.e., HOSS and LOSS) that were derived during the creation of the vertical scale. Statistics from the simulations are organized in three general areas: bias of the ability estimates, magnitude of standard errors, and reliability.

Statistics for simulations that focused on ability estimation included the following:

- Bias: This is the statistical bias of the estimated theta parameter. This is a test of the assumption that error is randomly distributed around true ability. It is a measure of whether scores systematically underestimate or overestimate ability.
- Mean squared error (MSE): This is a measure of the magnitude of difference between true and estimated theta.
- Root mean squared error (RMSE): This is the square root of the MSE.
- Significance of the bias: This is an indicator of the statistical significance of bias.
- Average standard error of the estimated theta: This is the average of the simulated standard error of measurement. It is the marginal reliability for the simulated population.
- Standard error of theta at the 5th, 25th, 75th, and 95th percentiles
- Percentage of students' estimated theta falling outside the 95% and 99% confidence intervals.
- Reliability of ability estimates

The relevant computational details for the statistics used to summarize the simulations are provided below and are followed by summary tables for each full test configuration and testlet as well as some associated plots to facilitate the interpretability of the results.

Bias

At the test and reporting category levels, the bias is the difference between actual ability, θ_j, and estimated ability, $\widehat{\theta_j}$, for *j*th student.

$$Bias_j = \theta_j - \hat{\theta}_j ,$$

The average bias over examinees, $\overline{Bias}$, is defined as

$$\overline{Bias} = \frac{\sum_{j=1}^{N}(\theta_j - \hat{\theta}_j)}{N}.$$

N is the number of total students. The standard deviation of the estimated bias is

$$SD_{Bias} = \sqrt{\frac{\sum_{j=1}^{N}\left(\theta_j - \bar{\hat{\theta}}_j\right)^2}{N-1}},$$

where $\bar{\hat{\theta}}_j$ is the average of estimated ability, $\hat{\theta}_j$, and the standard error of the estimated bias is

$$SE_{Bias} = \frac{SD_{Bias}}{\sqrt{N}}$$

The average bias is tested for statistical significance with a z-test,

$$z_{Bias} = \frac{\overline{Bias}}{SE_{Bias}}.$$

The z-statistic follows standard normal distribution, N(0,1), and the *p*-value for a two-tailed test is reported. The mean squared error (MSE) is the average of squared bias,

$$MSE = \frac{\sum_{j=1}^{N}(\theta_j - \hat{\theta}_j)^2}{N}.$$

At a student level, the degree of the deviation of the estimated ability from the actual ability is assessed with z-test

$$z_{Bias_j} = \frac{\theta_j - \hat{\theta}_j}{SE_{\hat{\theta}_j}}$$

where $SE_{\hat{\theta}_j}$ is the standard error of the estimated score for *j*th student. The percentages of students who are outside of 95% and 99% of the confidence interval are computed and reported. The critical values used for the confidence interval are 1.96 and 2.58.

Tables 38 through 48 provide the estimated bias of the estimated proficiencies for scores produced within DRC BEACON when the full ELA and mathematics test configurations are administered and when different testlets are administered. The bias in the overall total scores tends to be small and non-significant. Not surprisingly, the bias increases when scores are based on the administration of fewer items such as reporting category scores or in the smaller testlet configurations. Note that Figures 4 through 15 plot the associated bias relative to estimated ability to provide a graphic illustration of the relationship when the full ELA and mathematics tests are administered.

Table 38. Summary of Bias ELA Full Test Total

Grade	Mean Bias	SE of Mean Bias	P-value Bias	MSE	95% CI Miss Rate	99% CI Miss Rate
3	-0.01	0.02	0.50	0.12	6.03	1.70
4	-0.01	0.02	0.69	0.12	7.53	1.97
5	-0.02	0.02	0.27	0.12	7.07	2.00
6	-0.01	0.02	0.66	0.13	7.50	1.77
7	-0.01	0.03	0.62	0.20	8.10	2.03
8	0.01	0.03	0.87	0.23	8.00	2.37

Table 39. Summary of Bias Mathematics Full Test Total

Grade	Mean Bias	SE of Mean Bias	P-value Bias	MSE	95% CI Miss Rate	99% CI Miss Rate
3	0.03	0.01	0.02	0.10	8.20	2.30
4	0.03	0.02	0.03	0.12	9.00	1.97
5	0.03	0.02	0.07	0.15	8.93	2.83
6	0.05	0.02	0.00	0.16	9.90	2.67
7	0.05	0.02	0.01	0.25	10.10	3.13
8	0.04	0.02	0.09	0.28	9.93	3.00

Table 40. Summary of Bias ELA Full Test Reporting Categories

Grade	Reporting Category	Mean Bias	SE of Mean Bias	P-value Bias	MSE	95% CI Miss Rate	99% CI Miss Rate
3	Total	-0.01	0.02	0.50	0.12	6.03	1.70
	Reading: Key Ideas and Details	0.09	0.03	0.00	1.32	2.33	0.27
	Reading: Craft Structure/Integration of Knowledge and Ideas	0.09	0.03	0.00	1.61	2.50	0.37
	Reading: Vocabulary Acquisition and Use	-0.02	0.03	0.45	1.08	2.07	0.27
	Writing - Text Types and Purposes	0.07	0.03	0.01	1.69	2.00	0.37
	Writing - Conventions of Standard English	0.10	0.03	0.00	1.83	1.90	0.20
	Writing - Research	0.21	0.03	0.00	1.90	2.17	0.37
	Listening	0.04	0.03	0.20	1.04	2.13	0.27
4	Total	-0.01	0.02	0.69	0.12	7.53	1.97
	Reading: Key Ideas and Details	0.06	0.03	0.03	0.99	2.60	0.47
	Reading: Craft Structure/Integration of Knowledge and Ideas	0.00	0.03	0.99	1.91	1.80	0.30
	Reading: Vocabulary Acquisition and Use	-0.08	0.03	0.01	1.02	2.33	0.37
	Writing - Text Types and Purposes	0.04	0.03	0.19	1.81	1.87	0.30

Grade	Reporting Category	Mean Bias	SE of Mean Bias	P-value Bias	MSE	95% CI Miss Rate	99% CI Miss Rate
	Writing - Conventions of Standard English	0.03	0.03	0.41	1.83	1.47	0.07
	Writing - Research	0.11	0.03	0.00	1.57	1.67	0.27
	Listening	0.03	0.03	0.33	1.16	2.97	0.33
5	Total	-0.02	0.02	0.27	0.12	7.07	2.00
	Reading: Key Ideas and Details	-0.04	0.03	0.19	1.00	2.90	0.50
	Reading: Craft Structure/Integration of Knowledge and Ideas	-0.03	0.03	0.39	1.55	2.17	0.27
	Reading: Vocabulary Acquisition and Use	-0.07	0.03	0.05	1.21	2.67	0.47
	Writing - Text Types and Purposes	0.07	0.03	0.02	1.60	2.17	0.20
	Writing - Conventions of Standard English	0.01	0.03	0.75	1.70	1.90	0.50
	Writing - Research	0.04	0.03	0.19	1.87	1.60	0.20
	Listening	0.02	0.03	0.53	1.23	2.17	0.33
6	Total	-0.01	0.02	0.66	0.13	7.50	1.77
	Reading: Key Ideas and Details	-0.09	0.03	0.00	1.52	2.63	0.47
	Reading: Craft Structure/Integration of Knowledge and Ideas	-0.03	0.03	0.36	1.73	2.23	0.27
	Reading: Vocabulary Acquisition and Use	-0.06	0.03	0.03	1.14	2.07	0.57
	Writing - Text Types and Purposes	0.03	0.03	0.38	2.01	1.77	0.10
	Writing - Conventions of Standard English	0.14	0.03	0.00	2.17	1.83	0.13
	Writing - Research	0.13	0.03	0.00	2.26	1.57	0.10
	Listening	0.01	0.03	0.87	1.54	2.30	0.20
7	Total	-0.01	0.03	0.62	0.20	8.10	2.03
	Reading: Key Ideas and Details	-0.17	0.03	0.00	2.31	2.90	0.67
	Reading: Craft Structure/Integration of Knowledge and Ideas	-0.07	0.03	0.05	2.43	2.20	0.60

Grade	Reporting Category	Mean Bias	SE of Mean Bias	P-value Bias	MSE	95% CI Miss Rate	99% CI Miss Rate
	Reading: Vocabulary Acquisition and Use	-0.09	0.04	0.03	1.67	2.53	0.63
	Writing - Text Types and Purposes	0.09	0.04	0.01	3.20	1.70	0.27
	Writing - Conventions of Standard English	0.12	0.04	0.00	3.86	1.57	0.23
	Writing - Research	0.05	0.03	0.13	2.39	1.63	0.27
	Listening	0.04	0.04	0.21	2.04	2.13	0.27
8	Total	0.01	0.03	0.87	0.23	8.00	2.37
	Reading: Key Ideas and Details	-0.16	0.04	0.00	2.50	2.87	0.53
	Reading: Craft Structure/Integration of Knowledge and Ideas	-0.02	0.04	0.55	2.20	2.03	0.20
	Reading: Vocabulary Acquisition and Use	-0.21	0.05	0.00	2.52	2.47	0.70
	Writing - Text Types and Purposes	0.07	0.04	0.08	3.82	1.80	0.07
	Writing - Conventions of Standard English	0.10	0.04	0.01	5.09	1.63	0.10
	Writing - Research	0.03	0.04	0.37	2.48	2.77	0.37
	Listening	0.02	0.04	0.54	2.57	2.57	0.37

Table 41. Summary of Bias Mathematics Full Test Reporting Categories

Grade	Reporting Category	Mean Bias	SE of Mean Bias	P-value Bias	MSE	95% CI Miss Rate	99% CI Miss Rate
3	Total	0.03	0.01	0.02	0.10	8.20	2.30
	Algebra	0.01	0.02	0.40	0.35	2.87	0.60
	Number & Quantity	0.09	0.02	0.00	0.51	3.07	0.43
	Measurement & Data	0.08	0.02	0.00	0.66	2.60	0.33
	Geometry	0.09	0.02	0.00	0.89	1.67	0.13
4	Total	0.03	0.02	0.03	0.12	9.00	1.97
	Algebra	0.00	0.02	0.98	0.46	2.30	0.27
	Number & Quantity	0.06	0.02	0.00	0.53	3.10	0.43
	Measurement & Data	0.01	0.02	0.60	0.70	2.07	0.20
	Geometry	0.05	0.02	0.01	0.94	2.30	0.13
5	Total	0.03	0.02	0.07	0.15	8.93	2.83
	Algebra	0.03	0.02	0.18	0.56	2.87	0.43
	Number & Quantity	0.13	0.02	0.00	1.11	3.07	0.57
	Measurement & Data	0.07	0.02	0.00	0.86	2.63	0.30
	Geometry	0.08	0.02	0.00	1.04	1.90	0.20
6	Total	0.05	0.02	0.00	0.16	9.90	2.67
	Algebra	0.06	0.02	0.00	0.72	2.47	0.23
	Number & Quantity	0.11	0.02	0.00	0.71	3.23	0.60
	Measurement & Data	0.03	0.02	0.18	1.10	2.47	0.30
	Geometry	0.15	0.02	0.00	1.23	1.50	0.23
7	Total	0.05	0.02	0.01	0.25	10.10	3.13
	Algebra	0.10	0.02	0.00	1.26	2.27	0.40
	Number & Quantity	0.12	0.03	0.00	0.97	3.30	0.43
	Measurement & Data	0.05	0.03	0.03	1.20	1.97	0.23
	Geometry	0.11	0.02	0.00	1.67	2.30	0.40
8	Total	0.04	0.02	0.09	0.28	9.93	3.00
	Algebra	0.07	0.03	0.02	1.49	2.00	0.27
	Number & Quantity	0.15	0.03	0.00	1.20	2.77	0.37
	Measurement & Data	0.05	0.03	0.06	1.52	1.67	0.33
	Geometry	0.10	0.03	0.00	1.71	2.10	0.10

Table 42. Summary of Bias ELA Reporting Category Reading and Writing Testlet

Reporting Category	Grade	Mean Bias	SE of Mean Bias	P-value Bias	MSE	95% CI Miss Rate	99% CI Miss Rate
Total	3	0.00	0.02	0.95	0.20	10.17	3.07
	4	-0.06	0.02	0.00	0.19	10.27	3.57
	5	-0.08	0.02	0.00	0.18	9.67	2.90
	6	-0.07	0.02	0.00	0.19	9.90	2.50
	7	-0.08	0.03	0.00	0.31	10.10	3.53
	8	-0.13	0.03	0.00	0.33	9.43	3.00
Reading: Key Ideas and Details	3	0.06	0.03	0.03	1.30	2.63	0.37
	4	0.01	0.03	0.64	0.84	2.97	0.50
	5	-0.04	0.03	0.16	1.14	1.80	0.50
	6	-0.04	0.03	0.23	1.44	2.60	0.37
	7	-0.13	0.03	0.00	2.42	2.40	0.30
	8	-0.16	0.04	0.00	2.53	2.53	0.53
Reading: Craft Structure/Integration of Knowledge and Ideas	3	0.10	0.03	0.00	1.59	2.60	0.23
	4	0.05	0.03	0.11	1.61	2.33	0.30
	5	-0.07	0.03	0.02	1.51	2.93	0.33
	6	-0.01	0.03	0.64	1.66	2.27	0.33
	7	-0.06	0.03	0.08	2.40	2.73	0.63
	8	-0.02	0.04	0.52	2.10	2.27	0.50
Reading: Vocabulary Acquisition and Use	3	-0.03	0.03	0.34	1.11	2.43	0.30
	4	-0.09	0.03	0.00	1.05	2.70	0.43
	5	-0.03	0.03	0.43	1.22	2.57	0.50
	6	-0.05	0.03	0.07	1.05	2.37	0.17
	7	-0.07	0.04	0.11	1.62	2.73	0.30
	8	-0.21	0.05	0.00	2.12	2.40	0.40
Writing - Text Types and Purposes	3	0.10	0.03	0.00	1.75	2.03	0.20
	4	0.09	0.03	0.00	1.84	1.93	0.30
	5	0.06	0.03	0.04	1.58	1.73	0.23
	6	0.04	0.03	0.16	2.00	1.63	0.10
	7	0.08	0.04	0.02	3.00	1.57	0.17
	8	0.06	0.04	0.14	4.33	1.63	0.10
Writing - Conventions of Standard English	3	0.10	0.03	0.00	1.96	1.93	0.23
	4	0.03	0.03	0.36	2.02	2.37	0.50
	5	0.01	0.03	0.78	1.66	1.87	0.20
	6	0.07	0.03	0.02	2.06	1.83	0.17
	7	0.11	0.04	0.00	3.87	1.40	0.23
	8	0.13	0.04	0.00	5.20	1.27	0.07

Reporting Category	Grade	Mean Bias	SE of Mean Bias	P-value Bias	MSE	95% CI Miss Rate	99% CI Miss Rate
Writing - Research	3	0.20	0.03	0.00	1.87	2.27	0.37
	4	0.12	0.03	0.00	1.59	2.27	0.40
	5	0.09	0.03	0.00	1.90	2.03	0.23
	6	0.16	0.03	0.00	2.57	2.07	0.30
	7	0.12	0.03	0.00	2.27	1.87	0.17
	8	0.01	0.04	0.85	2.26	1.87	0.27

Table 43. Summary of Bias ELA Reporting Category Reading Testlets

Reporting Category	Grade	Mean Bias	SE of Mean Bias	P-value Bias	MSE	95% CI Miss Rate	99% CI Miss Rate
Total	3	-0.25	0.02	0.00	0.59	17.93	7.20
	4	-0.32	0.02	0.00	0.60	18.90	7.63
	5	-0.33	0.02	0.00	0.64	16.17	5.73
	6	-0.28	0.02	0.00	0.60	16.57	5.57
	7	-0.46	0.03	0.00	1.10	17.27	6.10
	8	-0.43	0.03	0.00	1.03	16.00	5.00
Reading: Key Ideas and Details	3	0.10	0.03	0.00	1.17	2.73	0.43
	4	0.01	0.03	0.62	1.02	2.47	0.43
	5	-0.05	0.03	0.09	1.22	2.57	0.33
	6	-0.12	0.03	0.00	1.97	2.43	0.40
	7	-0.19	0.03	0.00	2.64	2.63	0.53
	8	-0.17	0.04	0.00	2.72	2.43	0.43
Reading: Craft Structure/Integration of Knowledge and Ideas	3	0.14	0.03	0.00	1.57	2.10	0.47
	4	0.06	0.03	0.03	1.78	1.80	0.30
	5	0.04	0.03	0.17	1.52	2.37	0.40
	6	0.02	0.03	0.47	1.54	2.40	0.17
	7	-0.09	0.03	0.01	2.64	2.50	0.27
	8	-0.07	0.04	0.03	2.38	2.83	0.40
Reading: Vocabulary Acquisition and Use	3	-0.02	0.03	0.42	1.03	2.30	0.17
	4	-0.06	0.03	0.05	1.00	2.50	0.63
	5	-0.06	0.03	0.06	1.32	2.03	0.43
	6	-0.06	0.03	0.04	1.12	2.53	0.50
	7	-0.09	0.04	0.03	1.86	3.03	0.70
	8	-0.21	0.05	0.00	2.47	2.73	0.47

Table 44. Summary of Bias ELA Reporting Category Writing – Text Types & Purposes Testlet

Reporting Category	Grade	Mean Bias	SE of Mean Bias	P-value Bias	MSE	95% CI Miss Rate	99% CI Miss Rate
Writing – Text Types and Purposes	3	0.07	0.03	0.01	1.35	3.90	0.80
	4	0.05	0.03	0.12	1.08	4.27	0.87
	5	-0.01	0.03	0.86	1.17	3.73	0.53
	6	0.00	0.03	0.92	1.16	4.07	0.53
	7	0.01	0.04	0.82	1.94	4.63	1.13
	8	-0.04	0.04	0.33	2.28	4.53	0.77

Table 45. Summary of Bias ELA Reporting Category Writing – Conventions of Standard English Testlet

Reporting Category	Grade	Mean Bias	SE of Mean Bias	P-value Bias	MSE	95% CI Miss Rate	99% CI Miss Rate
Writing – Conventions of Standard English	3	0.05	0.03	0.09	1.21	3.67	0.47
	4	-0.01	0.03	0.83	1.14	4.73	0.87
	5	-0.02	0.03	0.44	1.15	3.83	0.53
	6	0.02	0.03	0.51	1.32	4.43	0.70
	7	0.04	0.04	0.24	2.02	4.07	0.63
	8	0.00	0.04	0.97	3.04	3.67	0.57

Table 46. Summary of Bias ELA Reporting Category Writing – Research Testlet

Reporting Category	Grade	Mean Bias	SE of Mean Bias	P-value Bias	MSE	95% CI Miss Rate	99% CI Miss Rate
Writing – Research	3	0.18	0.03	0.00	1.18	3.37	0.57
	4	0.08	0.03	0.01	0.94	3.80	0.67
	5	0.01	0.03	0.78	1.05	4.67	0.87
	6	0.05	0.03	0.10	1.51	4.23	0.37
	7	0.06	0.03	0.06	1.35	4.10	0.63
	8	0.02	0.04	0.53	1.76	3.43	0.57

Table 47. Summary of Bias ELA Reporting Category Listening Testlet

Reporting Category	Grade	Mean Bias	SE of Mean Bias	P-value Bias	MSE	95% CI Miss Rate	99% CI Miss Rate
Listening	3	0.07	0.03	0.01	1.20	3.20	0.37
	4	-0.02	0.03	0.58	1.05	2.77	0.43
	5	0.00	0.03	0.97	1.38	2.90	0.53
	6	-0.03	0.03	0.25	1.43	2.60	0.43
	7	0.05	0.04	0.16	2.26	2.47	0.60
	8	0.11	0.04	0.00	2.90	2.33	0.27

Table 48. Summary of Bias Mathematics Reporting Category Testlets

Reporting Category	Grade	Mean Bias	SE of Mean Bias	P-value Bias	MSE	95% CI Miss Rate	99% CI Miss Rate
Algebra	3	0.01	0.02	0.43	0.27	4.03	0.73
	4	0.01	0.02	0.45	0.31	4.60	1.10
	5	0.03	0.02	0.20	0.44	4.97	1.10
	6	0.08	0.02	0.00	0.55	4.33	0.87
	7	0.13	0.02	0.00	1.03	4.60	1.20
	8	0.17	0.03	0.00	1.30	4.00	1.03
Number & Quantity	3	0.12	0.02	0.00	0.50	4.33	0.93
	4	0.09	0.02	0.00	0.46	4.73	1.23
	5	0.16	0.02	0.00	0.81	4.67	1.17
	6	0.14	0.02	0.00	0.76	4.97	1.33
	7	0.14	0.03	0.00	0.90	4.63	0.83
	8	0.19	0.03	0.00	1.15	3.97	0.87
Measurement & Data	3	0.10	0.02	0.00	0.57	4.37	0.73
	4	0.03	0.02	0.19	0.46	4.63	0.83
	5	0.05	0.02	0.02	0.58	4.20	0.63
	6	0.09	0.02	0.00	0.74	4.23	0.73
	7	0.10	0.03	0.00	0.99	4.17	0.87
	8	0.11	0.03	0.00	1.43	3.97	0.73
Geometry	3	0.14	0.02	0.00	0.85	3.70	0.77
	4	0.09	0.02	0.00	0.79	4.37	0.83
	5	0.11	0.02	0.00	0.99	3.73	1.00
	6	0.25	0.02	0.00	1.49	4.43	0.77
	7	0.26	0.03	0.00	1.74	3.90	0.67
	8	0.19	0.03	0.00	1.67	4.00	0.77

Figure 4. Conditional Bias Plot ELA Grade 3

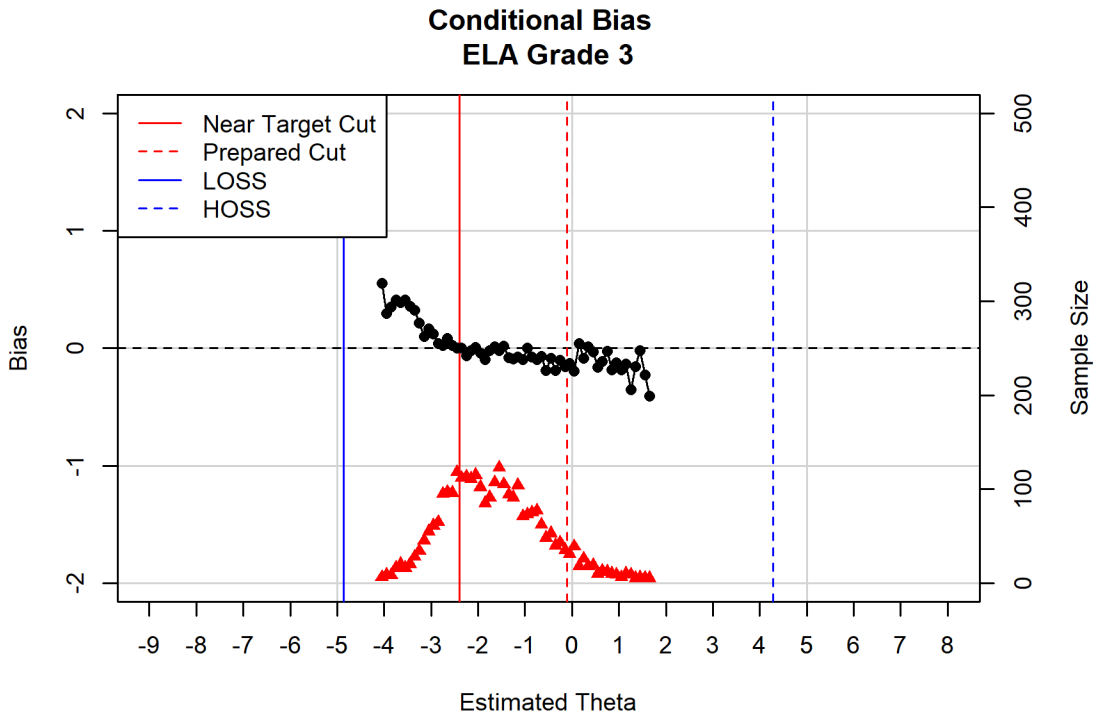


Figure 5. Conditional Bias Plot ELA Grade 4

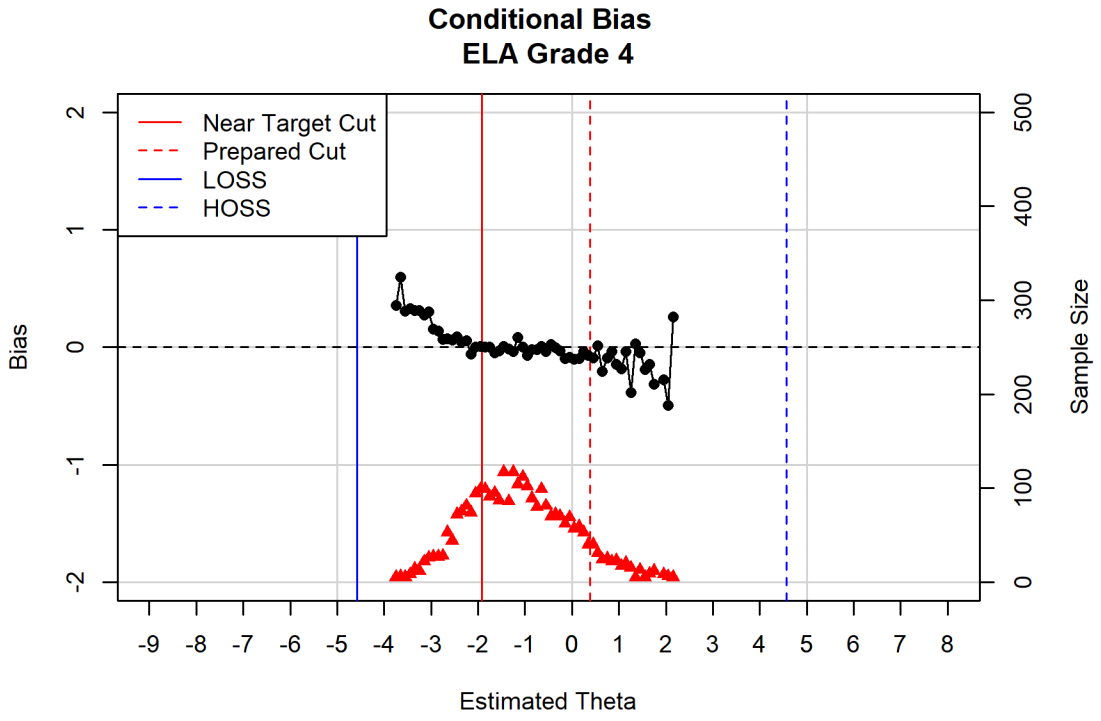


Figure 6. Conditional Bias Plot ELA Grade 5

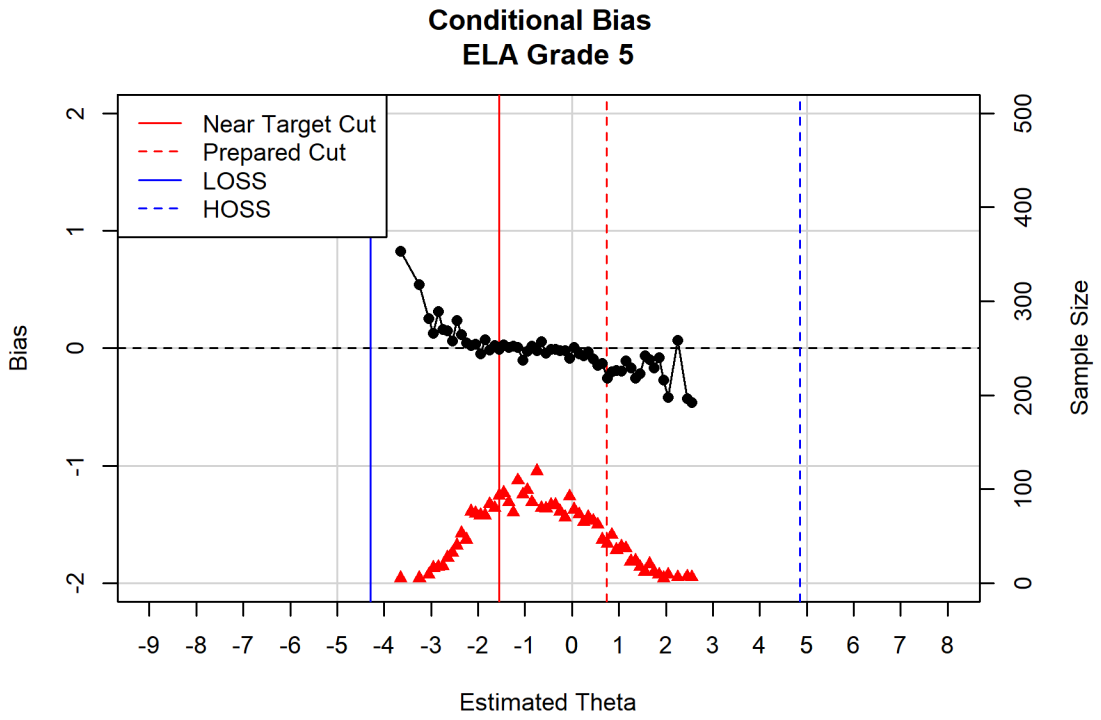


Figure 7. Conditional Bias Plot ELA Grade 6

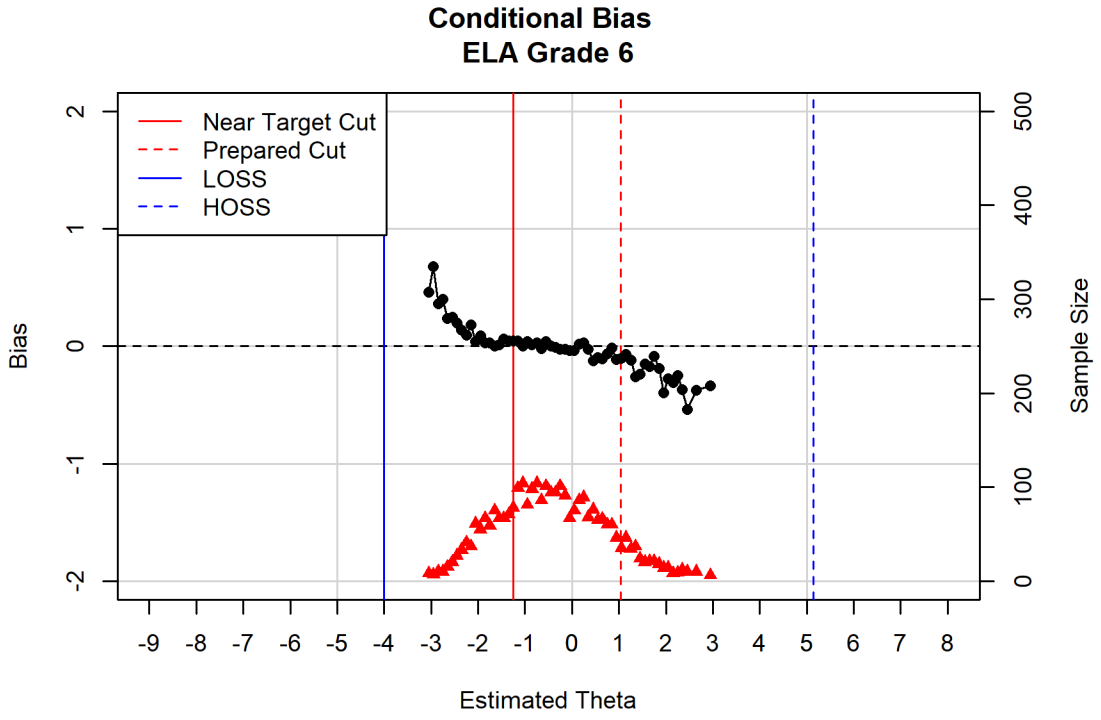


Figure 8. Conditional Bias Plot ELA Grade 7

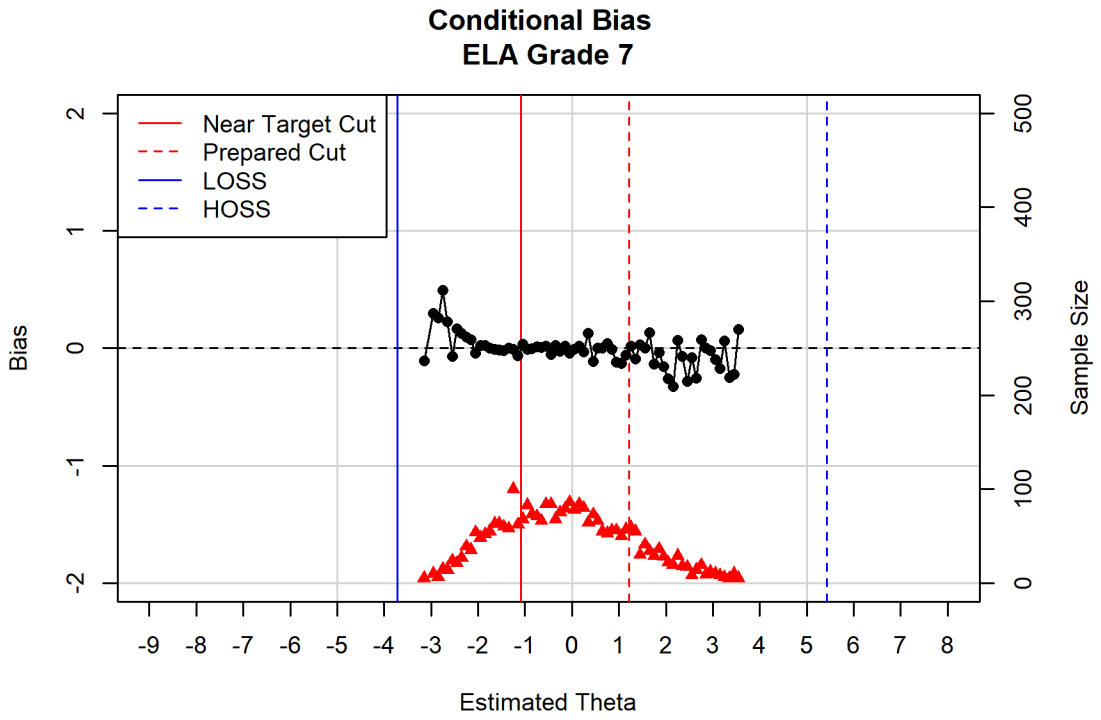


Figure 9. Conditional Bias Plot ELA Grade 8

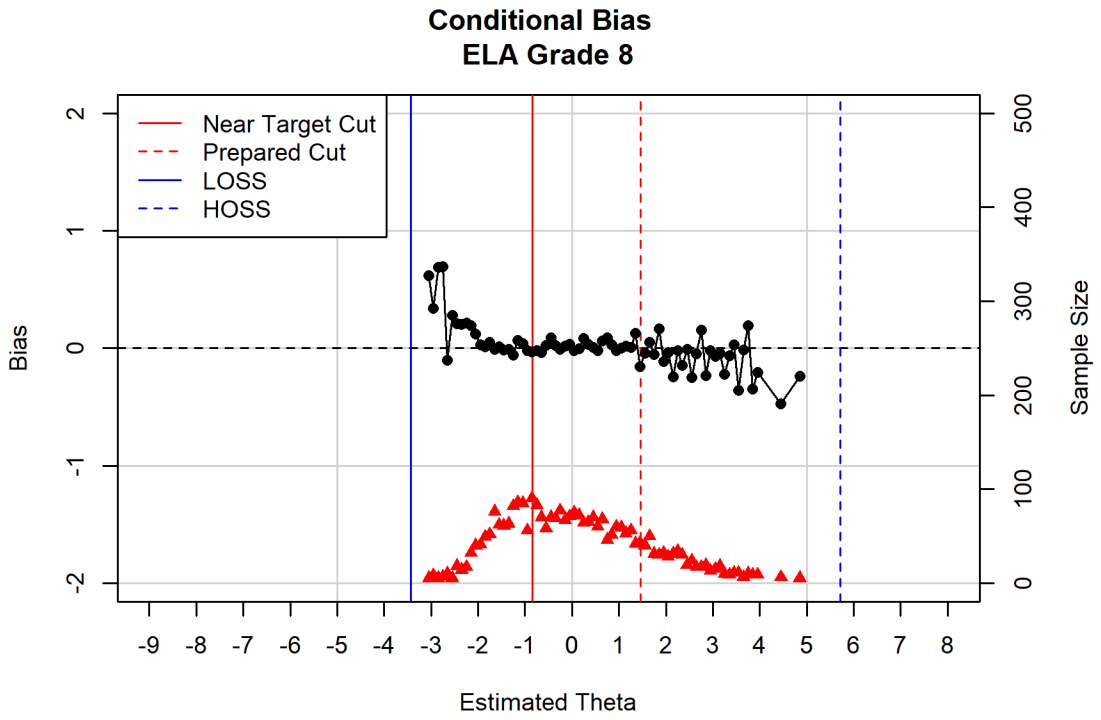


Figure 10. Conditional Bias Plot Mathematics Grade 3

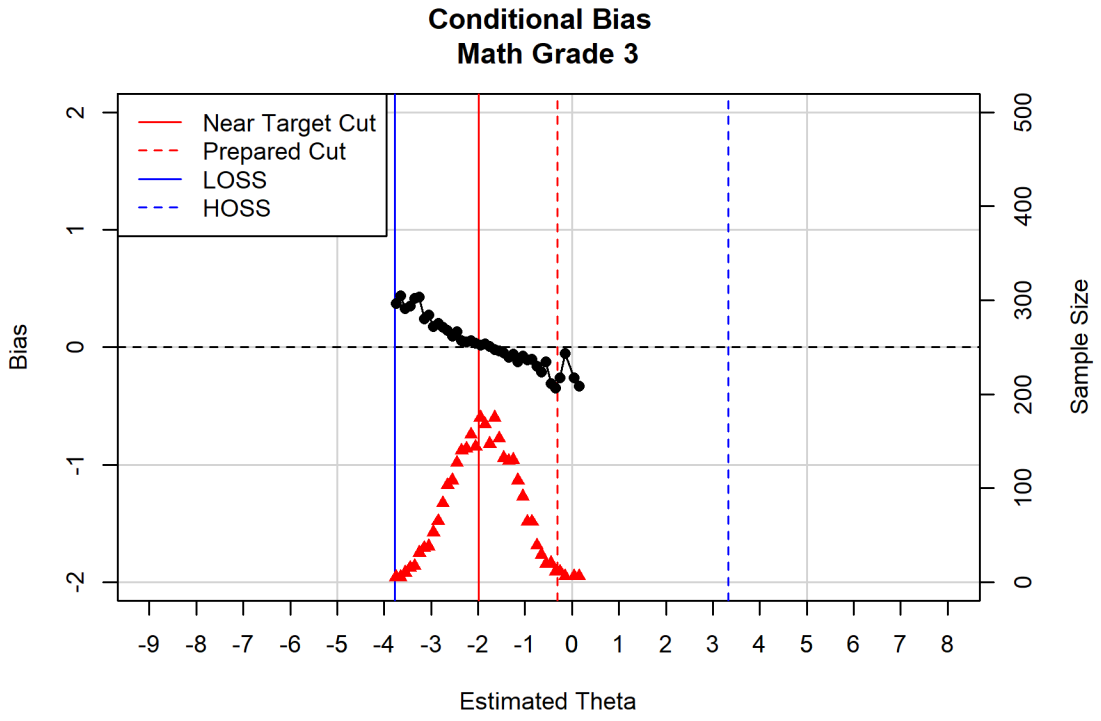


Figure 11. Conditional Bias Plot Mathematics Grade 4

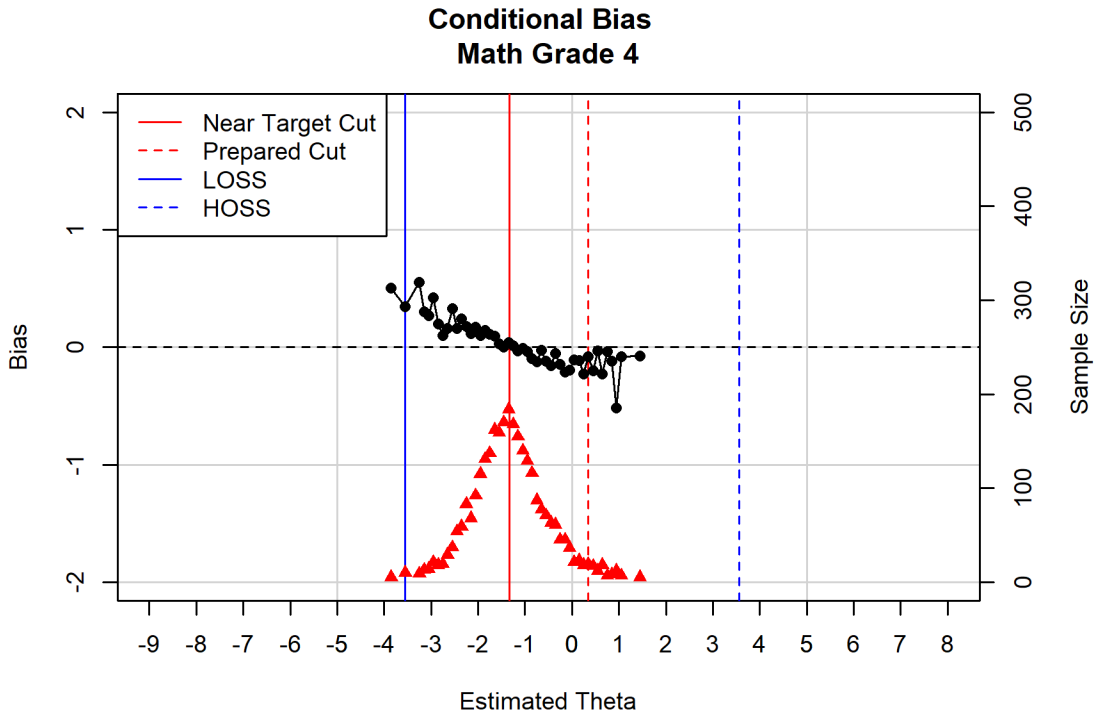


Figure 12. Conditional Bias Plot Mathematics Grade 5

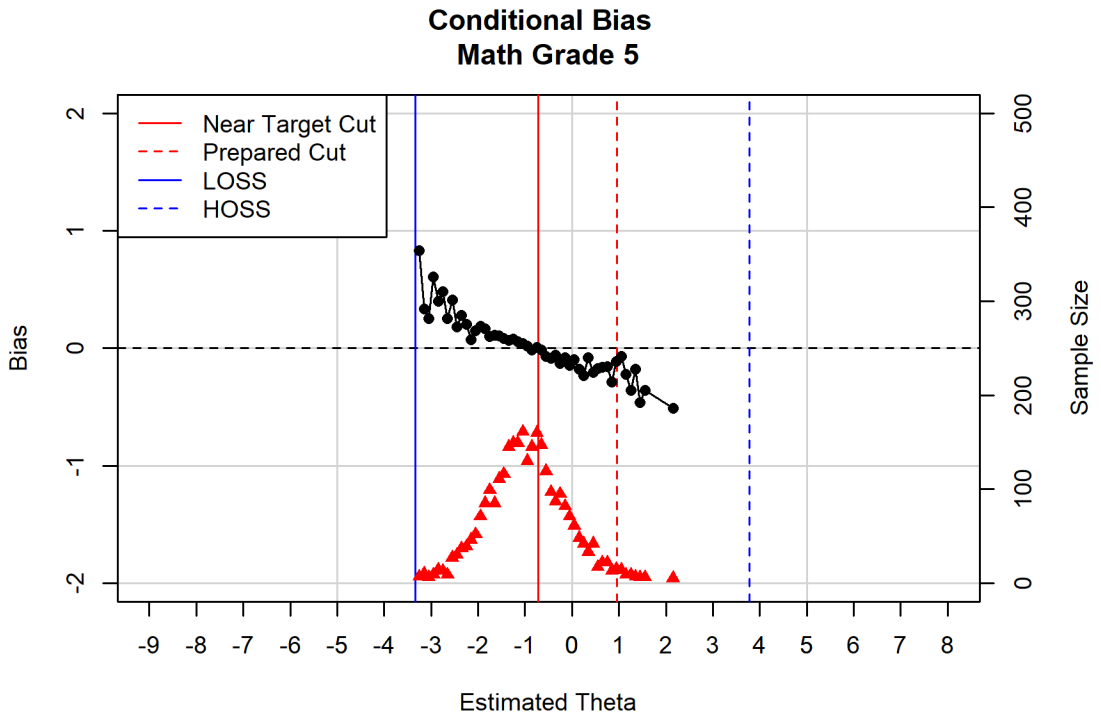


Figure 13. Conditional Bias Plot Mathematics Grade 6

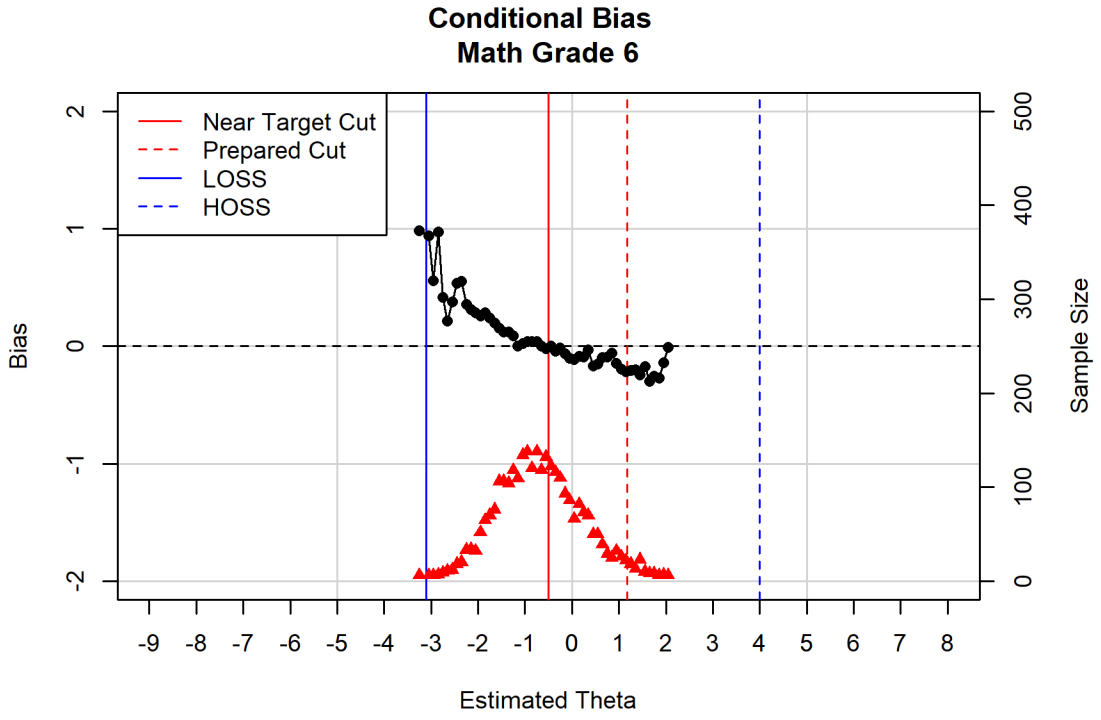


Figure 14. Conditional Bias Plot Mathematics Grade 7

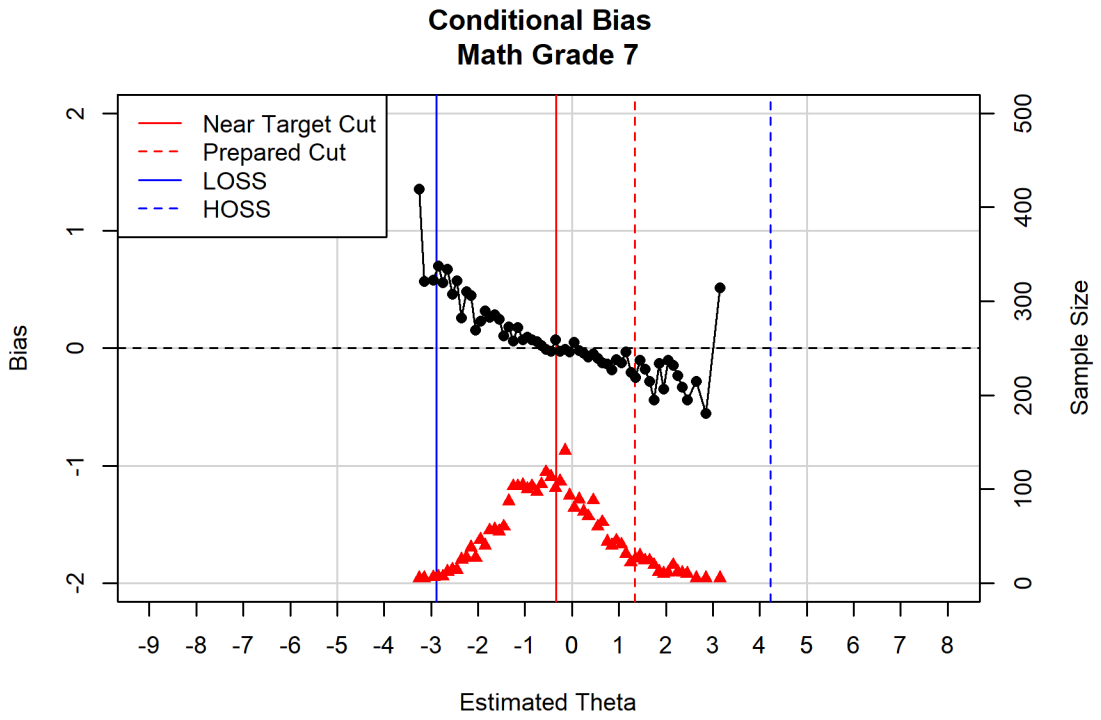
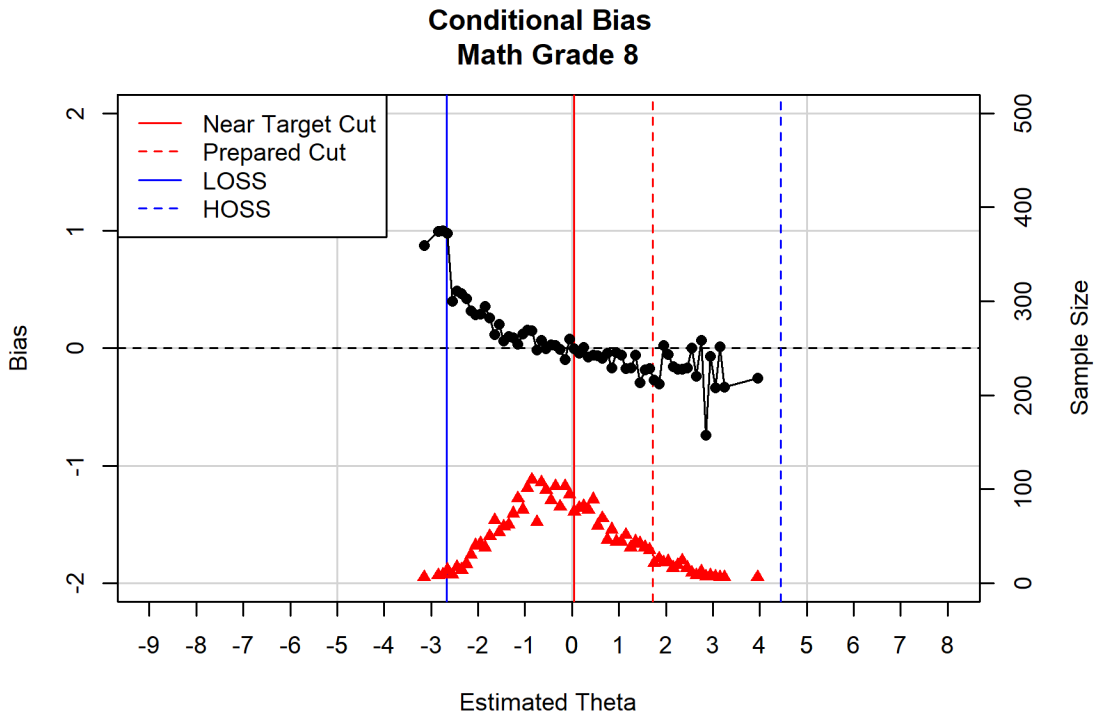


Figure 15. Conditional Bias Plot Mathematics Grade 8



Standard Error of Measurement (SEM)

The SEM associated with each ability estimate is calculated for each configuration of DRC BEACON using methods described in the previous chapter. Tables 49 through 52 provide statistical summaries (including the minimum, maximum, mean, median, and standard deviation) of the SEMs for mathematics and ELA. As the grade increases, the average SEM increases. This is possibly due to the mismatch between the item difficulty distributions and students' ability distributions in higher grades. The simulation consisted of actual students' abilities across the ability continuum. Tables 53 through 59 provide statistical summaries of the SEMs for testlets. Tables 60 through 70 summarize the SEM by deciles of estimated student performance in scale score units to highlight the error at associated points of the scale. Figures 16 through 27 provide graphical summaries of the SEM relative to estimated ability. As expected, the SEM estimates are higher for extreme scores in the highest and lowest deciles. Also, as expected, the SEM estimates are larger for scores based on fewer items such as reporting category scores and testlet scores.

Table 49. Summary of Standard Error of Measurement by Grade for Total ELA Full Tests

Level	Mean	Standard Deviation	Min	Max	Median
3	0.32	0.05	0.24	0.92	0.31
4	0.31	0.06	0.23	0.84	0.29
5	0.31	0.07	0.23	1.72	0.29
6	0.32	0.07	0.23	0.97	0.30
7	0.38	0.10	0.24	1.23	0.35
8	0.42	0.12	0.27	1.65	0.38

Table 50. Summary of Standard Error of Measurement by Grade for Total Mathematics Full Tests

Level	Mean	Standard Deviation	Min	Max	Median
3	0.26	0.07	0.20	1.74	0.24
4	0.28	0.07	0.22	1.47	0.27
5	0.31	0.08	0.23	2.46	0.29
6	0.32	0.08	0.25	1.62	0.30
7	0.37	0.12	0.28	4.75	0.35
8	0.43	0.13	0.31	5.22	0.40

Table 51. Summary of Standard Error of Measurement by Grade for Full ELA Tests

Grade	Reporting Category	Mean	Standard Deviation	Min	Max	Median
3	Total	0.32	0.05	0.24	0.92	0.31
	Reading: Key Ideas and Details	1.10	0.68	0.51	6.21	0.88
	Reading: Craft Structure/Integration of Knowledge and Ideas	1.23	0.80	0.55	6.17	0.94
	Reading: Vocabulary Acquisition and Use	1.09	0.51	0.51	3.05	0.91
	Writing – Text Types and Purposes	1.24	0.68	0.55	6.05	1.05
	Writing – Conventions of Standard English	1.32	0.69	0.64	6.43	1.09
	Writing – Research	1.35	0.80	0.59	6.30	1.05
	Listening	1.01	0.53	0.48	4.90	0.84
4	Total	0.31	0.06	0.23	0.84	0.29
	Reading: Key Ideas and Details	0.92	0.52	0.48	5.28	0.77
	Reading: Craft Structure/Integration of Knowledge and Ideas	1.22	0.89	0.56	8.49	0.92
	Reading: Vocabulary Acquisition and Use	1.09	0.58	0.47	6.61	0.87
	Writing – Text Types and Purposes	1.25	0.67	0.56	6.08	1.07
	Writing – Conventions of Standard English	1.32	0.70	0.61	6.04	1.09
	Writing – Research	1.25	0.68	0.58	7.07	1.02
	Listening	0.99	0.58	0.47	7.21	0.81
5	Total	0.31	0.07	0.23	1.72	0.29
	Reading: Key Ideas and Details	0.94	0.58	0.46	5.46	0.74
	Reading: Craft Structure/Integration of Knowledge and Ideas	1.13	0.82	0.50	8.68	0.85
	Reading: Vocabulary Acquisition and Use	1.15	0.67	0.52	6.53	0.91
	Writing – Text Types and Purposes	1.18	0.62	0.52	6.02	1.00
	Writing – Conventions of Standard English	1.23	0.64	0.62	6.11	1.03
	Writing – Research	1.26	0.79	0.53	7.76	1.01
	Listening	1.03	0.63	0.45	6.72	0.84
6	Total	0.32	0.07	0.23	0.97	0.30
	Reading: Key Ideas and Details	1.08	0.97	0.48	10.42	0.81
	Reading: Craft Structure/Integration of Knowledge and Ideas	1.17	0.87	0.55	7.40	0.89
	Reading: Vocabulary Acquisition and Use	1.07	0.54	0.52	5.67	0.88

Grade	Reporting Category	Mean	Standard Deviation	Min	Max	Median
	Writing – Text Types and Purposes	1.29	0.77	0.54	6.31	1.06
	Writing – Conventions of Standard English	1.35	0.85	0.62	6.96	1.08
	Writing – Research	1.37	0.89	0.52	6.88	1.09
	Listening	1.14	0.73	0.45	6.67	0.93
7	Total	0.38	0.10	0.24	1.23	0.35
	Reading: Key Ideas and Details	1.33	1.26	0.52	10.43	0.93
	Reading: Craft Structure/Integration of Knowledge and Ideas	1.40	1.10	0.60	10.69	1.04
	Reading: Vocabulary Acquisition and Use	1.33	0.61	0.60	4.78	1.15
	Writing – Text Types and Purposes	1.65	0.91	0.65	6.66	1.37
	Writing – Conventions of Standard English	1.74	1.17	0.71	8.55	1.32
	Writing – Research	1.43	0.85	0.54	6.58	1.17
	Listening	1.31	0.79	0.45	9.33	1.12
8	Total	0.42	0.12	0.27	1.65	0.38
	Reading: Key Ideas and Details	1.41	1.22	0.54	9.32	0.96
	Reading: Craft Structure/Integration of Knowledge and Ideas	1.38	0.99	0.65	9.94	1.06
	Reading: Vocabulary Acquisition and Use	1.68	1.03	0.60	5.64	1.28
	Writing – Text Types and Purposes	1.76	0.98	0.70	7.50	1.48
	Writing – Conventions of Standard English	2.02	1.27	0.77	8.16	1.56
	Writing – Research	1.49	0.81	0.62	6.79	1.25
	Listening	1.50	0.88	0.48	6.50	1.23

Table 52. Summary of Standard Error of Measurement by Grade for Full Mathematics Tests

Level	Reporting Category	Mean	Standard Deviation	Min	Max	Median
3	Total	0.26	0.07	0.20	1.74	0.24
	Algebra	0.58	0.28	0.33	4.13	0.48
	Number & Quantity	0.73	0.53	0.35	4.34	0.53
	Measurement & Data	0.75	0.47	0.36	4.12	0.59
	Geometry	0.87	0.61	0.48	6.37	0.67
4	Total	0.28	0.07	0.22	1.47	0.27
	Algebra	0.66	0.32	0.34	4.01	0.56
	Number & Quantity	0.76	0.45	0.39	5.84	0.59
	Measurement & Data	0.77	0.46	0.38	4.68	0.63
	Geometry	0.88	0.54	0.50	5.85	0.73
5	Total	0.31	0.08	0.23	2.46	0.29
	Algebra	0.74	0.39	0.37	3.93	0.62
	Number & Quantity	0.96	0.82	0.40	6.76	0.67
	Measurement & Data	0.84	0.53	0.39	5.83	0.67
	Geometry	0.94	0.57	0.52	6.76	0.77
6	Total	0.32	0.08	0.25	1.62	0.30
	Algebra	0.79	0.49	0.41	6.68	0.64
	Number & Quantity	0.83	0.60	0.38	6.45	0.61
	Measurement & Data	0.93	0.56	0.47	6.13	0.77
	Geometry	1.02	0.66	0.48	5.60	0.79
7	Total	0.37	0.12	0.28	4.75	0.35
	Algebra	1.00	0.67	0.49	6.90	0.79
	Number & Quantity	0.97	0.64	0.43	4.44	0.72
	Measurement & Data	1.02	0.54	0.52	6.03	0.85
	Geometry	1.17	0.73	0.53	9.71	0.93
8	Total	0.43	0.13	0.31	5.22	0.40
	Algebra	1.15	0.62	0.52	5.11	0.96
	Number & Quantity	1.14	0.75	0.44	5.08	0.86
	Measurement & Data	1.16	0.64	0.55	5.57	0.94
	Geometry	1.25	0.68	0.54	6.87	1.05

Table 53. Summary of Standard Error of Measurement ELA Reading and Writing Testlet

Reporting Category	Grade/Level	Mean	Standard Deviation	Min	Max	Median
Total	3	0.36	0.07	0.27	1.42	0.34
	4	0.35	0.07	0.25	1.95	0.33
	5	0.34	0.08	0.25	1.19	0.31
	6	0.35	0.09	0.24	1.43	0.33
	7	0.43	0.12	0.28	1.37	0.39
	8	0.46	0.14	0.29	1.67	0.41
Reading: Key Ideas and Details	3	1.07	0.62	0.51	5.92	0.88
	4	0.87	0.44	0.44	5.09	0.75
	5	0.96	0.63	0.45	5.73	0.75
	6	1.04	0.86	0.48	10.00	0.81
	7	1.36	1.18	0.53	9.69	0.97
	8	1.41	1.09	0.54	8.70	1.02
Reading: Craft Structure/Integration of Knowledge and Ideas	3	1.21	0.77	0.57	7.77	0.95
	4	1.20	0.81	0.52	7.16	0.95
	5	1.09	0.76	0.51	7.88	0.81
	6	1.14	0.79	0.52	7.12	0.88
	7	1.35	1.01	0.63	9.94	1.02
	8	1.30	0.94	0.62	9.72	1.02
Reading: Vocabulary Acquisition and Use	3	1.09	0.50	0.57	3.01	0.90
	4	1.11	0.61	0.47	6.40	0.87
	5	1.16	0.74	0.56	6.43	0.88
	6	1.05	0.55	0.52	5.40	0.86
	7	1.32	0.63	0.60	5.14	1.12
	8	1.63	1.02	0.60	5.31	1.21
Writing - Text Types and Purposes	3	1.25	0.67	0.51	5.07	1.06
	4	1.27	0.65	0.58	4.91	1.10
	5	1.18	0.64	0.51	5.41	0.99
	6	1.29	0.76	0.51	6.74	1.08
	7	1.64	0.86	0.62	6.08	1.37
	8	1.83	1.02	0.70	7.49	1.54
Writing - Conventions of Standard English	3	1.33	0.70	0.65	6.34	1.09
	4	1.35	0.71	0.62	5.55	1.12
	5	1.22	0.65	0.59	5.66	1.01
	6	1.32	0.85	0.62	7.69	1.07
	7	1.78	1.16	0.72	7.97	1.36
	8	2.06	1.26	0.77	8.14	1.63

Reporting Category	Grade/Level	Mean	Standard Deviation	Min	Max	Median
Writing - Research	3	1.34	0.79	0.62	5.89	1.05
	4	1.27	0.65	0.57	5.25	1.07
	5	1.28	0.79	0.56	7.01	1.03
	6	1.44	0.98	0.54	6.64	1.11
	7	1.45	0.79	0.58	5.84	1.21
	8	1.49	0.72	0.62	6.25	1.29

Table 54. Summary of Standard Error of Measurement ELA Reading Testlet

Reporting Category	Grade/Level	Mean	Standard Deviation	Min	Max	Median
Total	3	0.50	0.19	0.34	5.58	0.46
	4	0.47	0.21	0.32	5.64	0.43
	5	0.49	0.26	0.32	6.27	0.43
	6	0.50	0.22	0.32	5.63	0.45
	7	0.63	0.27	0.35	5.16	0.55
	8	0.65	0.30	0.36	4.32	0.56
Reading: Key Ideas and Details	3	1.03	0.59	0.48	5.70	0.84
	4	0.93	0.53	0.46	5.18	0.77
	5	0.99	0.72	0.46	10.65	0.75
	6	1.16	1.15	0.47	10.72	0.81
	7	1.40	1.33	0.49	10.43	0.98
	8	1.45	1.25	0.54	9.55	1.00
Reading: Craft Structure/Integration of Knowledge and Ideas	3	1.21	0.78	0.53	8.21	0.93
	4	1.21	0.86	0.53	7.16	0.92
	5	1.14	0.80	0.52	7.58	0.86
	6	1.13	0.79	0.53	7.42	0.90
	7	1.42	1.10	0.60	8.86	1.06
	8	1.41	1.09	0.63	10.19	1.05
Reading: Vocabulary Acquisition and Use	3	1.07	0.48	0.58	5.56	0.90
	4	1.05	0.55	0.49	6.66	0.85
	5	1.18	0.70	0.49	6.06	0.91
	6	1.09	0.57	0.54	5.50	0.89
	7	1.39	0.61	0.64	4.97	1.23
	8	1.69	1.05	0.58	5.40	1.25

Table 55. Summary of Standard Error of Measurement ELA Writing – Text Types and Purposes Testlet

Reporting Category	Level	Mean	Standard Deviation	Min	Max	Median
Writing – Text Types and Purposes	3	0.94	0.60	0.50	5.74	0.75
	4	0.87	0.47	0.49	5.66	0.71
	5	0.86	0.58	0.48	6.18	0.69
	6	0.87	0.56	0.50	5.90	0.69
	7	1.05	0.63	0.54	6.06	0.84
	8	1.19	0.68	0.61	6.20	0.96

Table 56. Summary of Standard Error of Measurement ELA Writing – Conventions of Standard English Testlet

Reporting Category	Level	Mean	Standard Deviation	Min	Max	Median
Writing – Conventions of Standard English	3	0.91	0.55	0.54	6.83	0.74
	4	0.90	0.48	0.53	5.67	0.74
	5	0.88	0.46	0.57	5.19	0.75
	6	0.91	0.56	0.57	6.42	0.75
	7	1.08	0.71	0.60	6.53	0.85
	8	1.31	0.90	0.65	6.97	0.97

Table 57. Summary of Standard Error of Measurement ELA Writing – Research Testlet

Reporting Category	Level	Mean	Standard Deviation	Min	Max	Median
Writing – Research	3	0.98	0.66	0.52	6.54	0.74
	4	0.88	0.52	0.50	6.72	0.71
	5	0.82	0.56	0.45	6.54	0.66
	6	0.93	0.67	0.46	5.75	0.73
	7	0.96	0.60	0.47	5.76	0.77
	8	1.10	0.71	0.52	6.01	0.87

Table 58. Summary of Standard Error of Measurement ELA Listening Testlet

Reporting Category	Level	Mean	Standard Deviation	Min	Max	Median
Listening	3	1.03	0.56	0.48	4.90	0.84
	4	0.98	0.56	0.47	5.96	0.77
	5	1.05	0.74	0.46	7.18	0.82
	6	1.14	0.82	0.45	5.99	0.83
	7	1.36	0.89	0.47	8.52	1.08
	8	1.57	0.97	0.52	7.54	1.23

Table 59. Summary of Standard Error of Measurement for Mathematics Testlets

Reporting Category	Level	Mean	Standard Deviation	Min	Max	Median
Algebra	3	0.50	0.25	0.31	3.67	0.42
	4	0.51	0.22	0.31	3.48	0.46
	5	0.59	0.32	0.36	4.75	0.50
	6	0.64	0.40	0.38	5.21	0.53
	7	0.77	0.52	0.46	6.33	0.63
	8	0.95	0.60	0.51	4.78	0.76
Number & Quantity	3	0.61	0.49	0.33	5.49	0.44
	4	0.59	0.42	0.35	5.55	0.46
	5	0.68	0.59	0.36	5.70	0.50
	6	0.69	0.59	0.36	5.83	0.50
	7	0.78	0.62	0.40	4.60	0.56
	8	0.89	0.71	0.41	4.78	0.63
Measurement & Data	3	0.61	0.45	0.31	5.20	0.49
	4	0.57	0.29	0.35	4.16	0.51
	5	0.64	0.38	0.40	4.78	0.55
	6	0.72	0.39	0.45	4.44	0.62
	7	0.82	0.46	0.48	5.12	0.68
	8	0.96	0.58	0.55	5.07	0.78
Geometry	3	0.76	0.60	0.45	6.36	0.59
	4	0.72	0.50	0.46	6.59	0.60
	5	0.77	0.56	0.45	6.62	0.63
	6	0.88	0.66	0.43	6.01	0.66
	7	0.99	0.73	0.45	6.07	0.75
	8	1.03	0.67	0.50	5.75	0.83

Table 60. Conditional SEMs by Student Decile Including LOSS/HOSS for ELA Full Test Totals

Level	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
3	27.85	23.33	21.46	20.24	19.97	20.37	20.87	21.55	22.92	28.08
4	24.54	19.88	18.75	18.82	19.38	19.95	20.60	21.85	24.08	29.86
5	23.18	18.81	17.90	18.03	18.52	19.22	20.59	22.22	24.49	31.33
6	24.18	19.04	18.15	18.43	19.30	20.70	22.03	23.51	26.09	32.82
7	28.34	21.54	20.32	20.88	22.57	24.49	26.16	28.95	32.59	41.72
8	28.73	22.99	22.11	22.64	24.48	26.48	28.41	31.53	35.73	47.46

Table 61. Conditional SEMs by Student Decile Including LOSS/HOSS for Mathematics Full Test Totals

Level	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
3	35.81	25.83	23.31	21.67	20.77	20.60	20.68	21.137	21.97	24.67
4	34.94	24.74	23.443	23.27	23.31	23.553	23.94	24.643	25.71	28.52
5	38.62	27.86	25.947	25.63	25.70	25.787	26.07	26.39	26.62	28.47
6	42.807	29.65	27.803	27.18	26.85	26.693	26.70	26.637	26.84	28.59
7	51.06	34.90	32.34	31.02	30.23	30.12	30.40	31.187	31.36	34.11
8	56.577	39.82	36.327	34.68	33.89	34.46	35.36	35.83	36.53	40.11

Table 62. Conditional SEMs by Student Decile Including LOSS/HOSS for ELA Full Test and Reporting Categories

Level	Reporting Category	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
3	Total	27.85	23.33	21.46	20.24	19.97	20.37	20.87	21.55	22.92	28.08
	Reading: Key Ideas and Details	179.9	83.18	63.24	57.31	55.01	53.41	55.09	58.43	63.74	103.43
	Reading: Craft Structure/Integration of Knowledge and Ideas	212.72	99.06	71.29	60.81	54.28	52.97	55.15	58.48	68.54	124.55
	Reading: Vocabulary Acquisition and Use	123.27	67.88	57.10	54.06	54.77	55.97	60.07	66.07	79.64	146.20
	Writing - Text Types and Purposes	171.79	88.85	71.50	63.38	62.95	61.56	63.98	70.42	78.78	133.34
	Writing - Conventions of Standard English	193.74	102.83	78.62	69.66	66.92	64.38	66.25	70.44	79.92	133.89
	Writing - Research	226.79	129.03	93.87	79.42	71.26	65.89	62.62	60.00	60.86	94.98
	Listening	138.90	73.54	59.78	53.65	51.67	50.91	51.87	54.55	62.05	110.43

Level	Reporting Category	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
4	Total	24.54	19.88	18.75	18.82	19.38	19.95	20.60	21.85	24.08	29.86
	Reading: Key Ideas and Details	126.3	61.20	52.02	48.48	47.98	48.73	50.37	53.22	58.39	100.37
	Reading: Craft Structure/Integration of Knowledge and Ideas	192.80	77.59	60.87	55.17	53.40	53.5	57.25	63.50	77.13	164.15
	Reading: Vocabulary Acquisition and Use	93.48	52.50	49.40	50.53	53.15	57.46	63.19	74.78	100.74	169.82
	Writing - Text Types and Purposes	169.08	84.33	71.51	63.75	65.07	63.84	68.12	74.32	81.77	134.48
	Writing - Conventions of Standard English	174.59	92.48	72.74	65.96	65.25	65.32	69.05	73.99	86.15	155.67
	Writing - Research	186.7	104.25	83.91	70.16	63.84	61.44	60.17	60.05	67.75	118.56
	Listening	125.34	66.78	55.41	51.34	49.38	49.89	52.08	55.83	64.93	122.03
5	Total	23.18	18.81	17.90	18.03	18.52	19.22	20.59	22.22	24.49	31.33
	Reading: Key Ideas and Details	111.36	55.20	47.7	45.09	44.54	46.72	49.01	53.25	64.86	139.49
	Reading: Craft Structure/Integration of Knowledge and Ideas	157.3	69.17	54.89	51.63	49.62	50.31	52.84	59.73	75.38	171.67
	Reading: Vocabulary Acquisition and Use	93.55	54.18	51.41	52.34	54.69	58.85	65.63	79.07	106.00	189.77
	Writing - Text Types and Purposes	164.43	84.99	66.87	61.24	60.73	62.15	63.77	67.41	76.50	119.79
	Writing - Conventions of Standard English	161.11	82.48	68.63	63.65	63.40	65.04	67.60	72.76	82.94	136.83
	Writing - Research	187.82	91.45	72.75	63.40	58.87	57.05	59.66	67.21	77.19	149.57
	Listening	136.40	67.13	54.57	51.49	50.20	51.26	54.54	58.14	68.42	128.12
6	Total	24.18	19.04	18.15	18.43	19.30	20.70	22.03	23.51	26.09	32.82
	Reading: Key Ideas and Details	114.86	59.25	50.16	47.30	47.32	49.35	53.14	59.05	74.61	201.46
	Reading: Craft Structure/Integration of Knowledge and Ideas	163	68.18	56.06	53.34	52.30	54.25	57.68	64.44	78.09	168.28
	Reading: Vocabulary Acquisition and Use	114.89	60.26	51.82	50.31	51.43	55.52	60.50	68.43	85.75	149.43

Level	Reporting Category	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
	Writing - Text Types and Purposes	187.66	88.69	68.94	63.58	61.88	64.60	68.04	71.93	83.40	146.96
	Writing - Conventions of Standard English	213.68	87.32	73.42	66.51	66.30	66.73	69.43	73.68	83.89	142.02
	Writing - Research	224.39	100.16	78.11	66.34	62.31	64.62	65.70	70.66	80.36	144.78
	Listening	152.3	71.47	55.98	52.87	53.28	55.45	58.59	63.77	77.67	153.63
7	Total	28.34	21.54	20.32	20.88	22.57	24.49	26.16	28.95	32.59	41.72
	Reading: Key Ideas and Details	133.88	60.84	52.98	50.71	53.31	57.33	64.40	74.61	98.29	284.94
	Reading: Craft Structure/Integration of Knowledge and Ideas	175.7	73.92	62.00	59.19	60.14	62.91	68.99	79.04	98.51	242.07
	Reading: Vocabulary Acquisition and Use	113.99	61.46	57.29	61.77	68.00	76.80	86.76	101.7	120.35	179.96
	Writing - Text Types and Purposes	237.7	119.48	91.51	83.68	81.50	78.82	85.69	92.80	107.07	176.15
	Writing - Conventions of Standard English	280.3	119.72	94.33	82.74	78.54	81.21	83.28	89.38	104.48	207.05
	Writing - Research	203.58	100.94	78.46	70.50	68.12	70.17	72.12	78.83	90.61	167.96
	Listening	175.80	87.29	67.55	62.68	60.98	63.65	68.78	75.36	90.72	165.22
8	Total	28.73	22.99	22.11	22.64	24.48	26.48	28.41	31.53	35.73	47.46
	Reading: Key Ideas and Details	144.7	63.45	54.26	52.07	55.41	59.25	67.47	80.29	112.44	296.88
	Reading: Craft Structure/Integration of Knowledge and Ideas	176.4	80.04	68.95	63.91	62.33	63.83	67.82	75.21	93.43	212.06
	Reading: Vocabulary Acquisition and Use	98.94	60.70	61.95	69.43	79.14	91.96	111.21	138.8	178.54	288.80
	Writing - Text Types and Purposes	227.59	109.68	91.01	84.89	81.74	84.58	95.81	106.18	132.89	220.37
	Writing - Conventions of Standard English	312.34	127.06	103.80	93.19	92.16	92.89	99.62	110.5	135.03	247.22
	Writing - Research	189.37	101.14	79.62	73.26	71.76	73.39	76.98	85.05	100.74	189.34
	Listening	202.3	106.63	78.12	72.83	68.75	73.24	73.38	83.40	99.83	192.46

Table 63. Conditional SEMs by Student Decile Including LOSS/HOSS for Mathematics Full Test and Reporting Categories

Level	Reporting Category	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
3	Total	35.81	25.83	23.31	21.67	20.77	20.60	20.68	21.14	21.97	24.67
	Algebra	105.62	57.66	44.05	39.09	38.11	38.22	38.94	41.60	46.50	68.85
	Number & Quantity	167.81	88.97	61.75	48.95	45.55	42.34	41.83	41.66	43.94	72.49
	Measurement & Data	161.93	74.75	60.23	52.55	47.58	45.02	45.08	46.87	52.53	90.08
	Geometry	200.70	84.77	66.22	59.22	56.16	54.27	54.70	56.06	59.85	92.64
4	Total	34.94	24.74	23.44	23.27	23.31	23.55	23.94	24.64	25.71	28.52
	Algebra	108.90	54.99	48.66	45.27	44.88	46.55	48.39	52.18	58.09	85.60
	Number & Quantity	153.50	94.00	67.32	54.76	50.68	49.89	48.23	48.31	50.07	69.64
	Measurement & Data	131.26	60.89	53.12	51.07	50.43	51.24	54.98	58.95	68.77	115.90
	Geometry	168.10	76.66	67.65	62.22	60.86	60.44	61.40	63.16	68.86	104.36
5	Total	38.62	27.86	25.95	25.63	25.70	25.79	26.07	26.39	26.62	28.47
	Algebra	137.17	69.50	55.95	50.50	49.81	50.90	53.38	55.29	59.92	86.11
	Number & Quantity	252.36	108.32	74.99	64.74	57.16	56.87	54.10	54.76	56.03	85.83
	Measurement & Data	166.21	70.29	59.83	57.25	56.56	55.32	56.70	59.13	66.11	104.58
	Geometry	187.85	89.73	74.21	67.22	65.35	62.19	64.18	64.48	70.22	104.89
6	Total	42.81	29.65	27.80	27.18	26.85	26.69	26.70	26.64	26.84	28.59
	Algebra	154.42	74.20	60.36	56.04	54.14	54.28	53.91	55.71	60.43	90.43
	Number & Quantity	194.73	99.98	67.14	54.78	52.29	51.74	50.93	50.53	51.42	74.44
	Measurement & Data	174.31	79.28	69.99	66.55	64.67	63.55	65.84	70.26	75.28	111.27
	Geometry	224.36	108.80	81.50	70.74	66.80	63.14	63.53	65.35	69.13	105.97
7	Total	51.06	34.90	32.34	31.02	30.23	30.12	30.40	31.19	31.36	34.11
	Algebra	207.09	93.81	75.50	64.86	63.63	64.36	66.87	69.61	74.73	118.00
	Number & Quantity	219.17	113.48	77.77	67.34	61.18	59.51	57.64	58.07	64.31	97.63
	Measurement & Data	181.43	88.73	77.32	71.90	72.19	70.79	73.91	77.09	81.20	127.33
	Geometry	240.75	123.10	92.93	78.98	76.76	73.87	76.72	78.39	82.27	128.33
8	Total	56.58	39.82	36.33	34.68	33.89	34.46	35.36	35.83	36.53	40.11
	Algebra	217.31	108.23	86.51	80.25	78.43	78.84	80.51	85.78	90.96	131.78
	Number & Quantity	261.55	145.23	92.34	79.21	72.43	67.29	65.79	67.26	72.53	101.85
	Measurement & Data	210.22	104.81	81.31	76.38	75.01	76.96	79.54	85.02	94.50	161.83
	Geometry	237.45	120.14	97.60	89.40	85.89	85.47	87.79	88.71	92.45	142.45

Table 64. Conditional SEMs by Student Decile Including LOSS/HOSS for ELA Reading and Writing Testlet

Reporting Category	Level	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
Total	3	32.43	25.69	23.50	22.14	21.75	22.24	23.07	23.78	25.47	31.25
	4	27.74	21.87	20.67	20.57	21.20	21.92	23.25	24.80	27.35	32.12
	5	25.93	20.89	19.49	19.89	20.30	21.02	22.02	24.05	26.92	35.93
	6	29.42	21.20	20.00	20.17	20.95	22.24	23.54	25.33	28.34	35.74
	7	32.55	24.11	22.50	23.64	25.41	26.79	28.54	31.71	36.78	47.37
	8	30.75	24.06	23.79	24.94	27.24	28.83	31.69	35.76	40.83	52.79
Reading: Key Ideas and Details	3	161.94	74.74	61.90	56.62	55.18	54.48	56.04	59.09	63.85	106.07
	4	115.67	59.25	50.69	46.86	46.58	48.02	50.04	52.76	55.91	85.58
	5	115.95	55.01	48.54	46.18	45.53	46.49	49.39	55.69	69.05	142.97
	6	118.24	59.17	50.60	48.26	48.53	49.94	52.69	58.00	71.61	171.98
	7	147.13	65.85	54.82	53.81	55.45	59.88	66.47	77.03	103.46	271.46
	8	153.89	65.60	57.22	55.53	59.71	64.41	71.57	86.15	115.24	259.95
Reading: Craft Structure/Integration of Knowledge and Ideas	3	205.60	94.45	69.96	60.22	55.42	53.85	54.74	59.36	69.38	122.57
	4	197.21	76.32	61.12	55.82	54.13	54.96	59.75	65.36	78.54	136.52
	5	142.22	61.95	52.55	49.77	48.05	49.01	51.89	58.46	75.23	172.40
	6	160.01	65.55	55.72	53.06	52.27	54.84	57.99	64.79	78.26	153.12
	7	166.38	70.64	62.17	58.99	59.93	63.85	69.32	78.54	97.02	221.35
	8	166.33	76.77	65.44	61.39	60.90	61.55	65.04	73.31	87.77	193.20
Reading: Vocabulary Acquisition and Use	3	120.62	67.51	56.25	54.69	55.33	56.41	59.85	65.50	79.74	146.84
	4	89.96	50.32	48.01	50.26	52.92	57.90	65.05	76.04	106.02	181.03
	5	95.10	54.24	52.09	51.79	54.21	57.95	64.19	76.29	103.40	201.52
	6	115.59	60.99	52.01	50.22	51.37	53.70	58.14	65.38	79.82	146.81
	7	115.65	62.07	57.45	59.90	65.99	74.33	83.57	97.50	116.88	188.76
	8	90.41	57.09	59.51	65.53	73.93	87.08	104.62	138.17	188.18	275.51
Writing - Text Types and Purposes	3	180.01	92.85	70.10	64.82	61.52	61.82	63.35	70.12	80.24	126.87
	4	175.48	88.70	73.51	65.85	65.73	65.77	69.58	73.60	82.00	128.82
	5	162.81	85.68	66.93	63.60	59.54	61.26	63.83	68.37	75.84	118.49
	6	181.08	89.12	69.82	63.73	61.28	64.61	66.53	73.34	84.63	146.84
	7	222.48	116.29	96.71	85.02	81.06	82.91	85.08	92.09	107.72	178.37
	8	243.08	109.26	96.65	83.05	83.74	86.77	98.96	116.40	145.82	216.11
Writing - Conventions of Standard English	3	189.41	102.43	78.33	71.89	64.51	65.19	66.14	70.89	80.24	140.20
	4	177.97	94.95	77.54	69.67	68.39	66.80	70.29	75.57	86.71	159.53
	5	153.96	79.48	67.61	62.23	62.46	63.36	66.70	71.22	80.98	146.15
	6	199.54	87.88	72.91	64.72	66.22	67.43	67.52	73.29	82.53	142.39
	7	282.31	121.36	95.91	84.13	81.33	82.20	84.81	95.75	109.49	211.41
	8	316.45	132.34	109.05	96.08	94.86	96.44	102.43	115.88	143.36	237.53

Reporting Category	Level	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
Writing - Research	3	225.08	127.17	94.08	79.26	73.10	65.96	61.74	58.79	60.89	95.12
	4	182.83	109.25	87.73	74.24	65.86	64.20	59.87	64.05	70.33	111.88
	5	197.44	90.57	73.86	65.56	61.14	60.76	61.43	66.53	76.12	140.76
	6	250.74	116.30	77.55	68.33	66.04	64.68	66.51	73.12	82.34	139.92
	7	214.06	107.41	82.83	77.11	70.64	71.57	74.34	79.58	91.45	148.47
	8	182.77	102.82	81.21	76.76	75.31	76.34	81.47	85.90	99.65	177.40

Table 65. Conditional SEMs by Student Decile Including LOSS/HOSS for ELA Reading Testlet

Reporting Category	Level	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
Total	3	52.93	35.61	32.15	30.20	29.65	29.86	30.63	31.99	34.00	44.98
	4	43.15	29.01	27.30	27.33	28.28	28.92	29.99	31.84	34.75	49.53
	5	42.35	29.81	27.93	27.30	27.26	27.78	29.54	32.41	37.69	61.90
	6	45.60	30.26	27.87	27.39	28.16	29.93	31.79	34.82	39.41	57.79
	7	48.18	33.18	31.23	32.01	34.80	37.64	40.89	46.02	53.34	80.30
	8	45.50	33.59	32.44	33.23	35.26	38.18	41.44	47.19	56.11	89.05
Reading: Key Ideas and Details	3	160.83	75.55	60.18	54.95	53.41	52.60	53.24	55.39	60.26	96.12
	4	124.43	63.90	53.55	49.16	48.43	48.45	49.31	51.49	56.90	103.51
	5	122.05	57.38	48.61	45.82	44.81	45.87	48.12	54.26	66.65	157.33
	6	128.35	59.01	49.60	46.85	48.15	49.14	52.90	60.94	78.37	239.87
	7	145.87	64.73	55.50	53.02	54.99	58.21	66.25	76.84	102.97	303.67
	8	153.17	66.30	57.40	55.73	58.26	62.18	68.81	82.08	112.29	299.65
Reading: Craft Structure/Integration of Knowledge and Ideas	3	215.70	99.84	72.56	62.53	56.25	53.76	53.25	57.17	65.69	109.84
	4	201.90	82.82	62.18	55.55	52.58	53.92	56.76	61.79	73.77	145.54
	5	166.22	72.95	56.90	52.98	51.12	52.12	53.03	59.33	72.06	160.85
	6	153.11	67.21	57.19	54.00	53.65	56.12	59.17	64.52	76.57	152.16
	7	172.14	72.42	64.42	60.55	62.65	64.93	70.75	79.43	101.24	247.64
	8	179.42	76.37	64.61	61.53	61.91	63.76	69.72	78.28	98.27	231.36
Reading: Vocabulary Acquisition and Use	3	121.24	69.70	59.45	55.23	55.16	56.18	57.76	63.34	75.41	135.49
	4	97.37	50.59	49.17	51.06	52.92	55.38	59.18	68.34	89.63	162.22
	5	104.89	56.41	52.92	51.98	53.69	58.22	66.48	78.08	107.28	197.88
	6	125.07	61.64	52.36	51.25	52.45	55.60	60.97	68.27	83.29	153.32
	7	128.17	70.09	62.16	65.58	73.14	82.41	91.61	101.93	117.63	179.00
	8	114.71	63.34	62.89	67.59	75.32	87.39	104.64	133.74	181.28	295.55

Table 66. Conditional SEMs by Student Decile Including LOSS/HOSS for ELA Writing – Text Types and Purposes Testlet

Reporting Category	Level	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
Writing – Text Types and Purposes	3	152.87	65.43	51.59	47.62	44.93	45.38	48.26	51.80	56.83	94.00
	4	121.99	58.00	48.42	44.14	42.75	44.42	46.50	50.54	57.80	93.37
	5	122.75	57.44	45.40	40.56	41.19	42.73	45.84	48.60	54.18	100.63
	6	117.93	56.37	44.04	41.02	42.52	44.54	46.95	49.07	55.00	109.04
	7	139.44	73.90	58.95	51.45	48.81	50.54	54.16	56.98	68.34	133.10
	8	132.03	73.06	62.29	57.34	56.52	58.12	62.15	69.70	92.77	170.37

Table 67. Conditional SEMs by Student Decile Including LOSS/HOSS for ELA Writing – Conventions of Standard English Testlet

Reporting Category	Level	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
Writing – Conventions of Standard English	3	123.96	63.17	54.17	49.41	46.73	45.73	47.69	49.96	53.83	99.51
	4	110.79	59.18	49.34	45.38	45.55	47.21	49.16	53.34	61.10	106.04
	5	99.17	54.73	47.82	46.00	46.66	48.47	50.59	54.65	61.27	105.82
	6	117.49	55.57	48.24	47.22	47.86	49.53	51.40	54.01	60.07	102.78
	7	156.36	71.45	57.68	53.72	52.74	53.41	54.58	58.64	66.12	130.78
	8	192.52	80.93	65.80	58.87	57.38	59.09	62.83	69.07	87.56	185.98

Table 68. Conditional SEMs by Student Decile Including LOSS/HOSS for ELA Writing – Research Testlet

Reporting Category	Level	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
Writing – Research	3	177.81	89.93	65.49	56.99	51.88	48.78	45.49	42.47	41.98	67.88
	4	134.06	66.80	53.98	49.95	47.29	44.80	42.80	43.19	50.02	81.35
	5	113.00	53.54	47.99	43.48	40.21	38.89	39.88	44.53	52.21	100.59
	6	149.02	61.10	50.62	44.89	42.80	43.54	47.11	51.10	59.23	104.17
	7	140.34	65.31	50.49	44.14	43.48	45.34	50.24	55.14	63.85	113.71
	8	151.82	71.30	53.89	47.80	48.81	52.83	57.34	63.39	75.64	149.51

Table 69. Conditional SEMs by Student Decile Including LOSS/HOSS for ELA Listening Testlet

Reporting Category	Level	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
Listening	3	149.45	72.94	61.383	54.74	52.52	51.79	51.91	53.77	61.47	110.74
	4	120.00	58.28	50.16	47.09	47.10	47.207	50.58	56.71	69.12	138.71
	5	151.24	66.08	50.073	46.19	45.28	48.43	52.36	58.967	71.53	145.86
	6	182.51	73.67	51.81	44.26	45.08	47.747	53.17	59.063	74.61	163.98
	7	199.18	96.12	69.957	60.27	58.54	62.333	65.83	73.197	87.46	181.79
	8	247.85	115.29	86.66	71.33	69.77	70.843	73.31	79.483	96.78	185.38

Table 70. Conditional SEMs by Student Decile Including LOSS/HOSS for Mathematics Testlets

Reporting Category	Level	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
Algebra	3	98.52	48.81	38.67	35.10	33.68	33.41	34.21	36.35	39.65	50.06
	4	82.63	41.92	37.77	37.19	37.44	38.57	40.18	42.96	46.29	55.50
	5	113.47	50.74	44.83	42.46	42.94	44.07	45.03	46.05	45.42	54.95
	6	131.30	59.55	49.88	46.22	45.89	46.38	46.17	45.70	46.63	56.74
	7	171.79	75.95	60.60	54.13	52.68	52.38	53.71	55.62	55.79	64.24
	8	212.18	96.90	72.53	65.59	63.84	64.85	65.68	66.17	66.70	81.77
Number & Quantity	3	160.39	72.01	49.89	41.47	38.03	36.91	36.04	36.20	36.69	44.00
	4	138.29	59.98	45.77	40.79	39.73	39.78	39.42	38.62	38.22	48.47
	5	185.64	70.49	51.65	45.21	43.51	43.26	43.22	42.40	41.86	48.35
	6	188.54	73.13	49.37	45.65	44.44	44.24	42.72	40.64	40.04	53.23
	7	208.06	81.74	55.96	49.56	48.90	47.37	46.41	46.31	48.82	65.49
	8	245.91	103.37	64.44	55.92	52.83	50.74	51.03	51.89	57.04	72.20
Measurement & Data	3	141.28	64.34	51.84	43.84	40.00	37.74	37.15	38.04	42.77	55.49
	4	96.14	50.88	43.17	41.56	41.69	43.05	44.78	45.21	47.98	61.94
	5	120.43	53.06	48.43	47.07	48.17	48.06	48.67	49.30	51.81	59.31
	6	138.29	63.20	56.32	54.18	53.51	52.79	54.62	55.72	56.74	62.36
	7	159.52	73.42	62.44	60.15	58.31	59.23	59.63	60.73	60.18	80.48
	8	201.12	87.20	69.40	66.14	65.71	66.92	67.58	67.80	69.30	103.83
Geometry	3	190.99	76.66	60.32	52.87	49.74	48.00	47.93	48.52	49.97	60.14
	4	154.57	63.89	55.51	52.42	51.54	51.55	51.69	53.45	54.46	61.88
	5	171.45	66.70	58.98	56.39	55.15	54.09	54.18	54.55	56.12	62.94
	6	219.74	98.32	71.92	62.45	56.59	54.67	55.13	55.33	54.30	60.57
	7	243.27	109.67	82.19	69.49	63.74	61.36	61.95	60.41	59.75	77.12
	8	231.18	109.64	83.23	73.60	67.45	70.17	68.93	68.17	66.26	92.11

Figure 16. Conditional SEM Plot ELA Grade 3

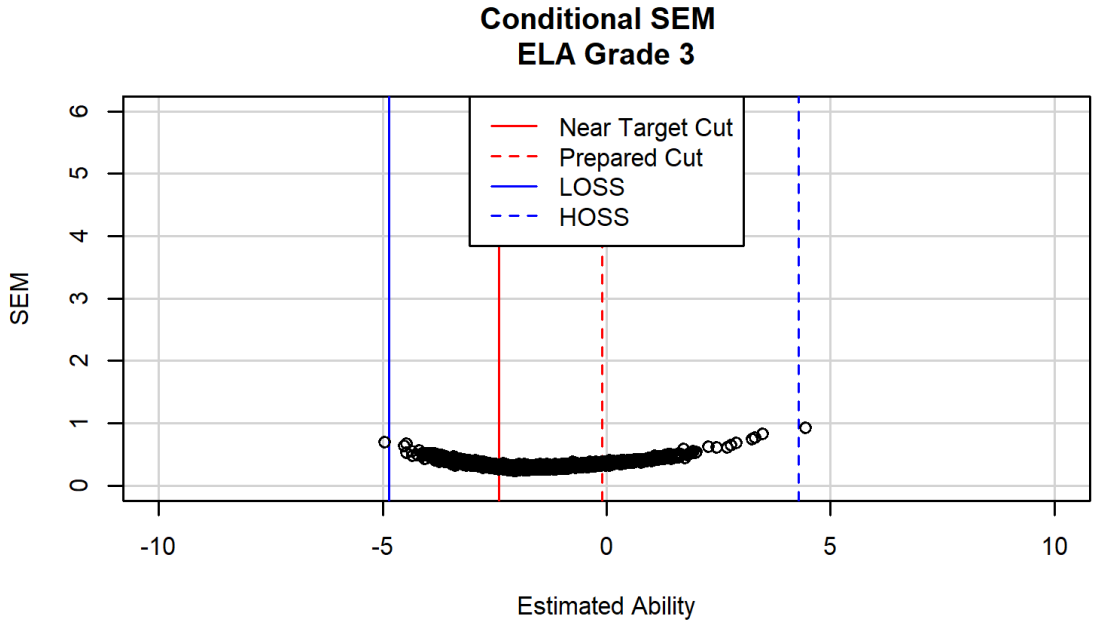


Figure 17. Conditional SEM Plot ELA Grade 4

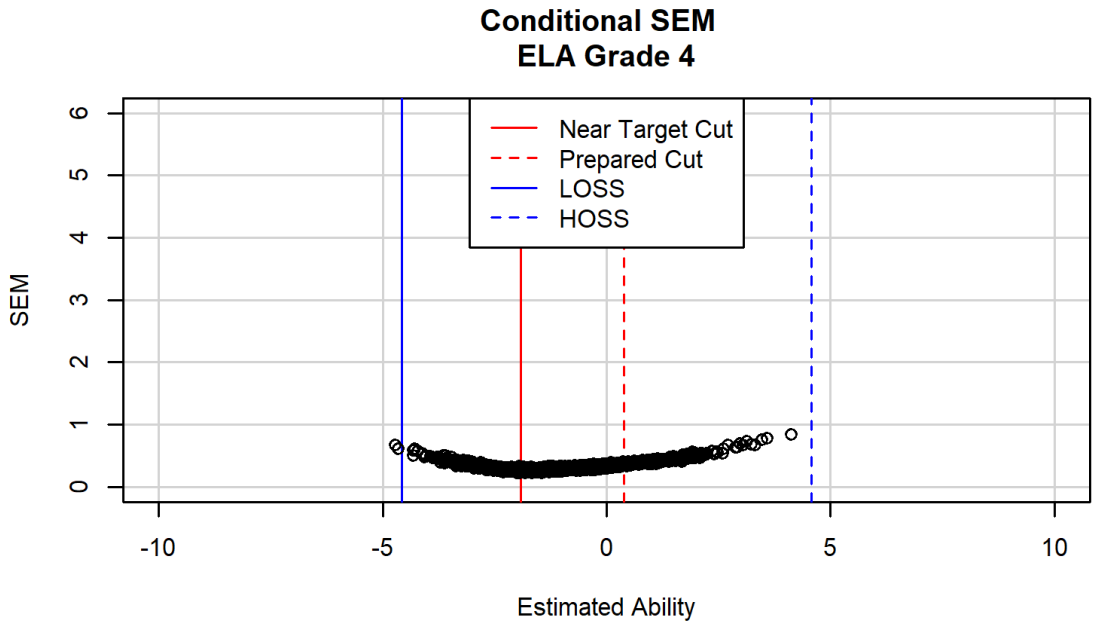


Figure 18. Conditional SEM Plot ELA Grade 5

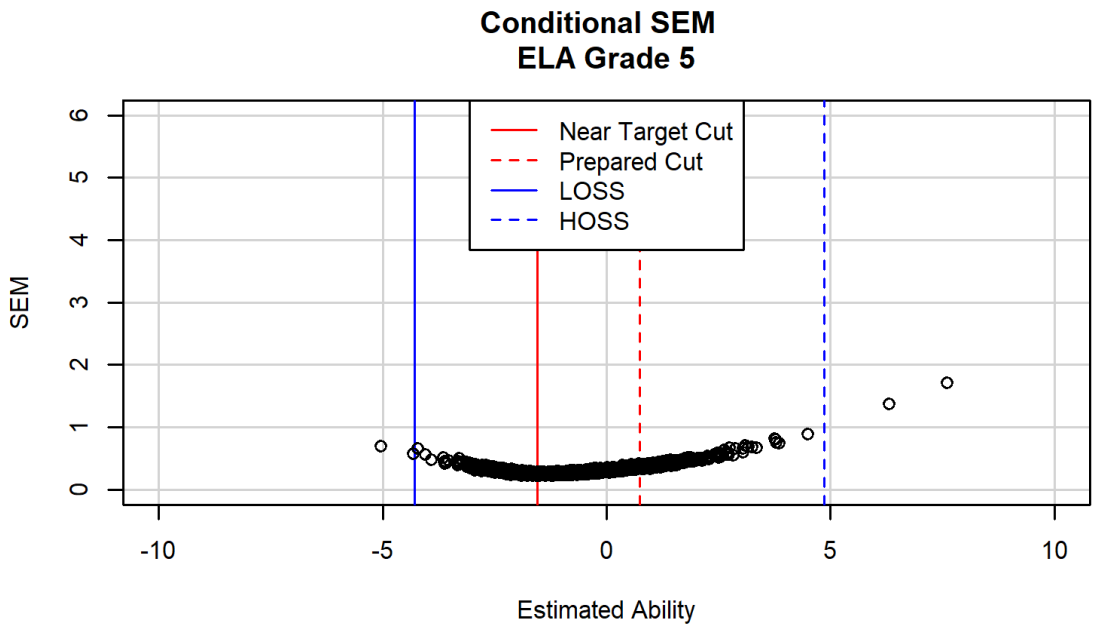


Figure 19. Conditional SEM Plot ELA Grade 6

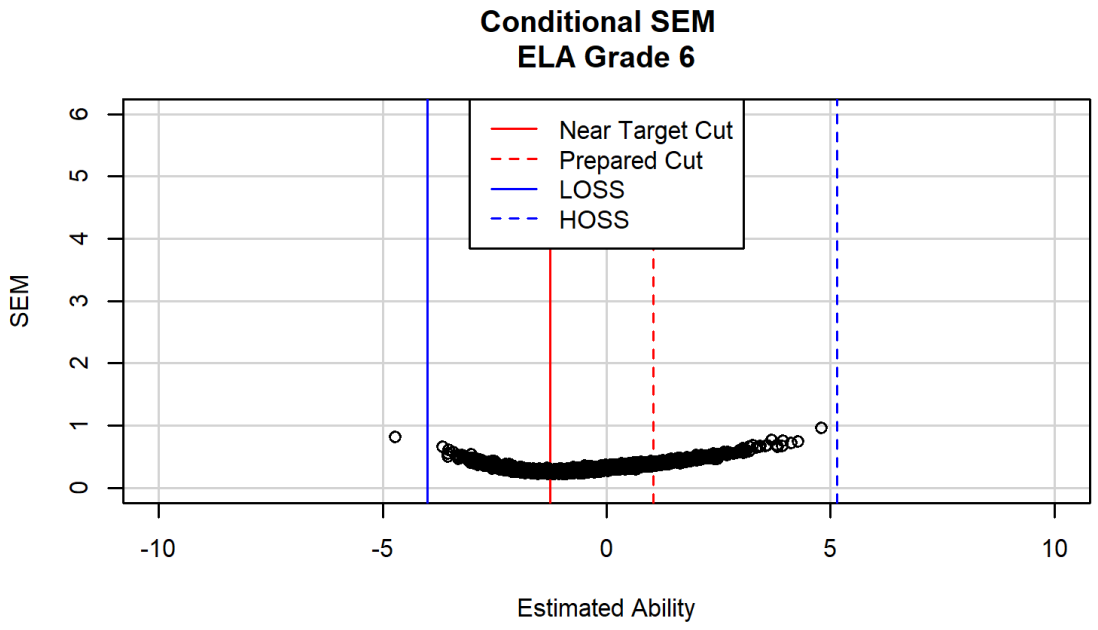


Figure 20. Conditional SEM Plot ELA Grade 7

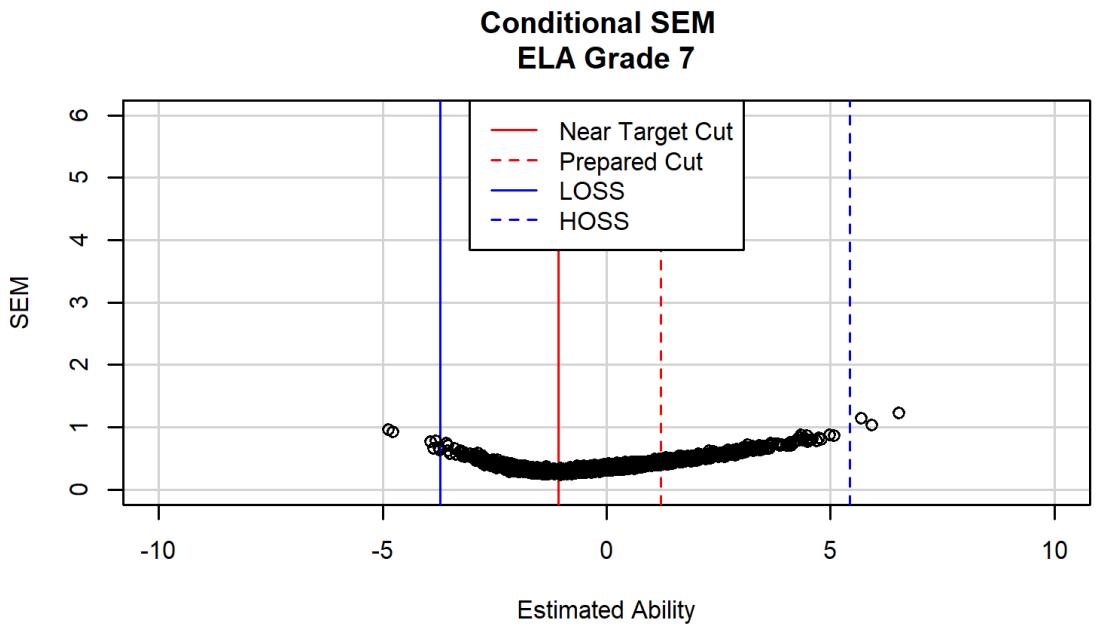


Figure 21. Conditional SEM Plot ELA Grade 8

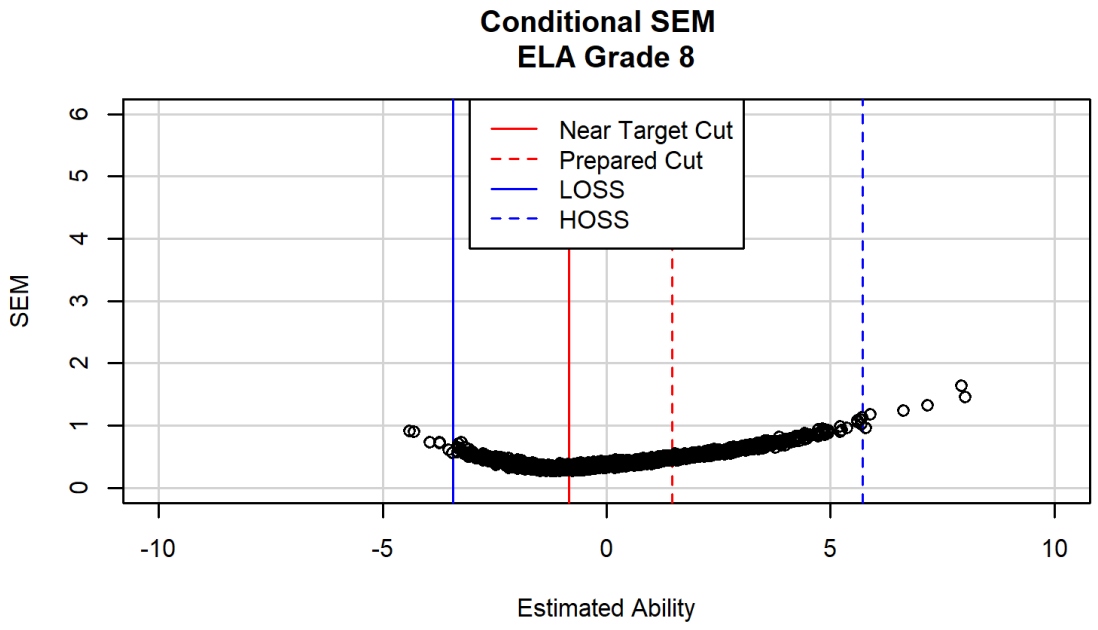


Figure 22. Conditional SEM Plot Mathematics Grade 3

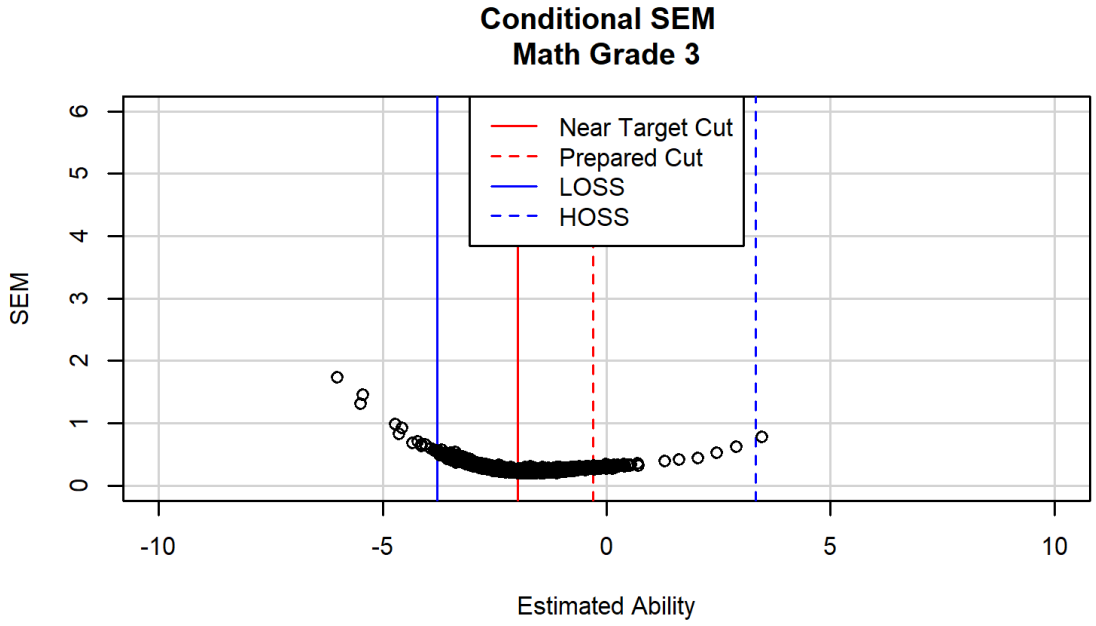


Figure 23. Conditional SEM Plot Mathematics Grade 4

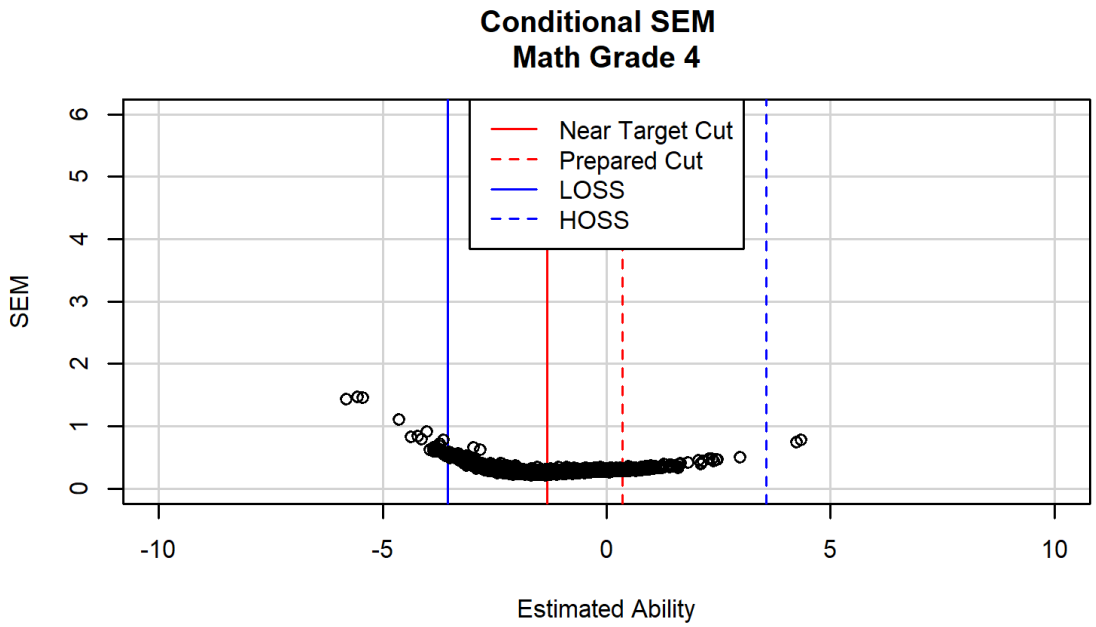


Figure 24. Conditional SEM Plot Mathematics Grade 5

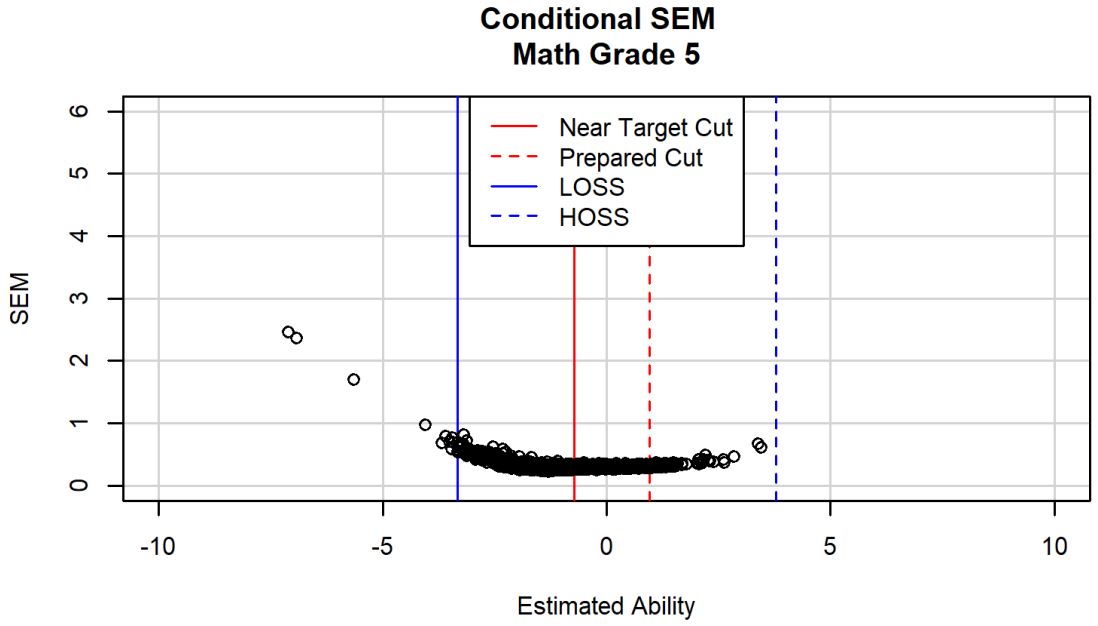


Figure 25. Conditional SEM Plot Mathematics Grade 6

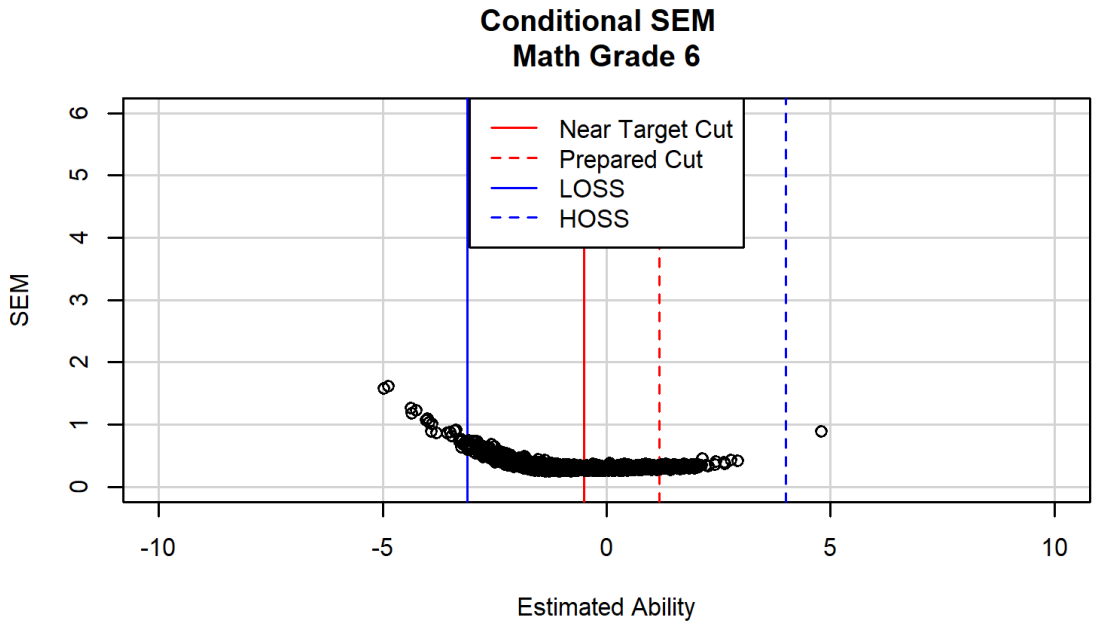


Figure 26. Conditional SEM Plot Mathematics Grade 7

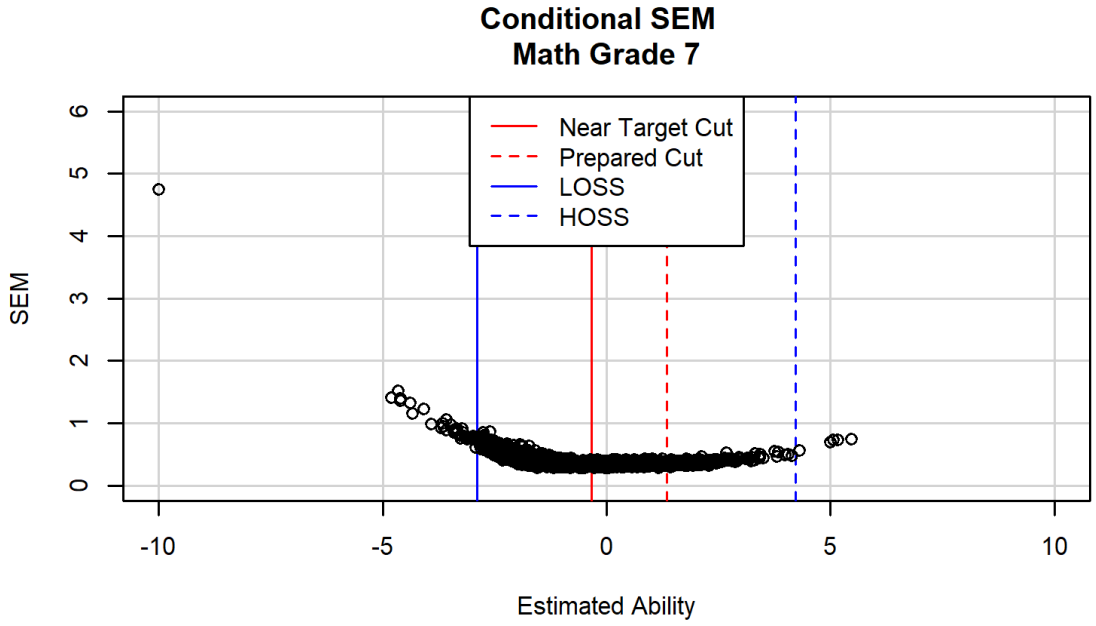
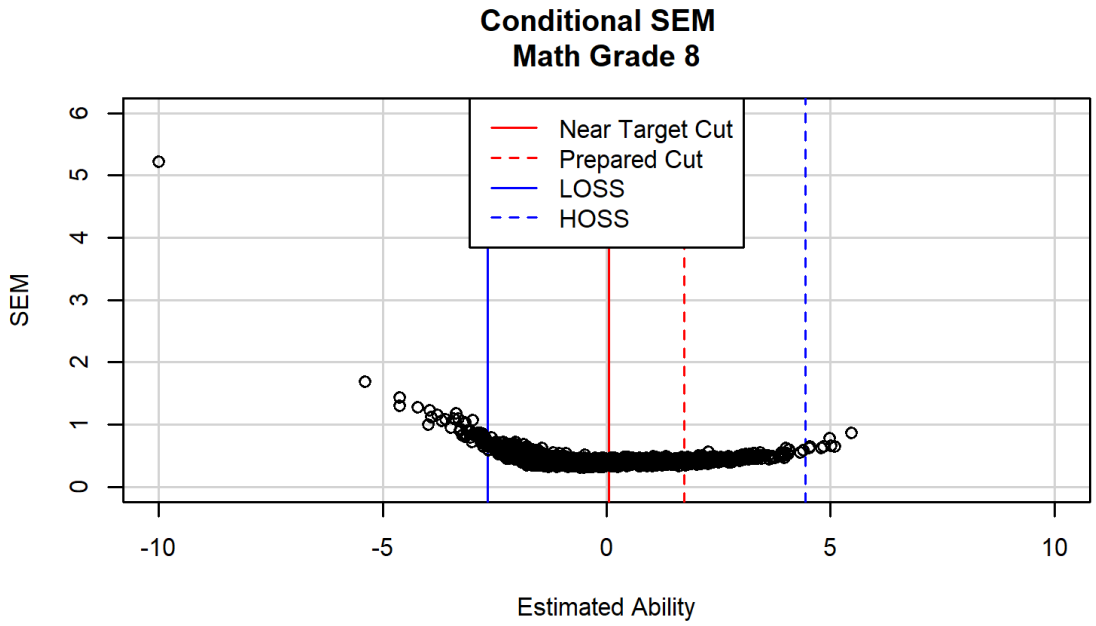


Figure 27. Conditional SEM Plot Mathematics Grade 8



Reliability

Reliability estimates reported in this tech report with simulated data are obtained with the following equation.

$$Reliability = \frac{MSE}{var(\hat{\theta})},$$

where $var(\hat{\theta})$ is the variance of the estimated ability. As explained in the Bias section, the mean squared error (MSE) is the average of squared bias. Root mean squared error (RMSE) is the square root of MSE.

Tables 71 through 81 show reliability estimates and precision for overall scores and associated reporting category scores for total tests and testlets. As expected, the overall estimated reliability is high and in the acceptable range for an interim assessment. Not surprisingly, the reliability estimates are lower for scores based on fewer items such as reporting category scores or testlet scores.

Table 71. Score Reliability of ELA Total Full Test

Level	Mean # Items	Mean SEM	Reliability	RMSE	SD Theta
3	59.90	0.32	0.90	0.35	1.12
4	59.70	0.31	0.91	0.35	1.17
5	59.70	0.31	0.92	0.35	1.21
6	59.70	0.32	0.91	0.37	1.23
7	59.70	0.38	0.91	0.45	1.48
8	59.70	0.42	0.91	0.48	1.61

Table 72. Score Reliability of Mathematics Total Full Test

Level	Mean # Items	Mean SEM	Reliability	RMSE	SD Theta
3	32	0.26	0.82	0.32	0.75
4	32	0.28	0.85	0.35	0.89
5	32	0.31	0.81	0.39	0.91
6	32	0.32	0.83	0.41	0.99
7	32	0.37	0.83	0.50	1.20
8	32	0.43	0.85	0.53	1.35

Table 73. Score Reliability of ELA Total Full Test and Reporting Categories

Grade	Reporting Category	Mean # Items	Mean SEM	Reliability	RMSE	SD Theta
3	Total	59.9	0.32	0.90	0.35	1.12
	Reading: Key Ideas and Details	9.9	1.10	0.69	1.15	2.07
	Reading: Craft Structure/Integration of Knowledge and Ideas	9.2	1.23	0.65	1.27	2.15
	Reading: Vocabulary Acquisition and Use	8.0	1.09	0.71	1.04	1.94
	Writing - Text Types and Purposes	8.0	1.24	0.63	1.30	2.14
	Writing - Conventions of Standard English	8.0	1.32	0.66	1.35	2.31
	Writing - Research	8.0	1.35	0.59	1.38	2.14
	Listening	8.8	1.01	0.74	1.02	2.00
4	Total	59.7	0.31	0.91	0.35	1.17
	Reading: Key Ideas and Details	9.8	0.92	0.71	1.00	1.84
	Reading: Craft Structure/Integration of Knowledge and Ideas	9.4	1.22	0.64	1.38	2.31
	Reading: Vocabulary Acquisition and Use	8.0	1.09	0.73	1.01	1.92
	Writing - Text Types and Purposes	8.0	1.25	0.63	1.35	2.21
	Writing - Conventions of Standard English	8.0	1.32	0.68	1.35	2.40
	Writing - Research	8.0	1.25	0.68	1.25	2.20
	Listening	8.5	0.99	0.71	1.08	2.00
5	Total	59.7	0.31	0.92	0.35	1.21
	Reading: Key Ideas and Details	9.8	0.94	0.74	1.00	1.97
	Reading: Craft Structure/Integration of Knowledge and Ideas	9.5	1.13	0.68	1.25	2.20
	Reading: Vocabulary Acquisition and Use	8.0	1.15	0.73	1.10	2.11
	Writing - Text Types and Purposes	8.0	1.18	0.63	1.26	2.08
	Writing - Conventions of Standard English	8.0	1.24	0.66	1.30	2.25
	Writing - Research	8.0	1.26	0.65	1.37	2.31
	Listening	8.5	1.03	0.70	1.11	2.04
6	Total	59.7	0.32	0.91	0.37	1.23
	Reading: Key Ideas and Details	9.6	1.08	0.69	1.23	2.22
	Reading: Craft Structure/Integration of Knowledge and Ideas	9.6	1.17	0.67	1.31	2.28
	Reading: Vocabulary Acquisition and Use	8.0	1.07	0.73	1.07	2.06
	Writing - Text Types and Purposes	8.0	1.29	0.64	1.42	2.36
	Writing - Conventions of Standard English	8.0	1.35	0.63	1.47	2.41
	Writing - Research	8.0	1.37	0.60	1.50	2.37
	Listening	8.4	1.14	0.70	1.24	2.27

Grade	Reporting Category	Mean # Items	Mean SEM	Reliability	RMSE	SD Theta
7	Total	59.7	0.38	0.91	0.45	1.48
	Reading: Key Ideas and Details	9.6	1.33	0.68	1.52	2.69
	Reading: Craft Structure/Integration of Knowledge and Ideas	9.7	1.40	0.67	1.56	2.71
	Reading: Vocabulary Acquisition and Use	8.0	1.33	0.78	1.29	2.73
	Writing - Text Types and Purposes	8.0	1.65	0.60	1.79	2.82
	Writing - Conventions of Standard English	8.0	1.74	0.58	1.97	3.05
	Writing - Research	8.0	1.43	0.66	1.54	2.63
	Listening	8.5	1.31	0.69	1.43	2.57
8	Total	59.7	0.42	0.91	0.48	1.61
	Reading: Key Ideas and Details	9.7	1.41	0.70	1.58	2.86
	Reading: Craft Structure/Integration of Knowledge and Ideas	9.5	1.38	0.69	1.48	2.65
	Reading: Vocabulary Acquisition and Use	8.0	1.69	0.72	1.59	3.00
	Writing - Text Types and Purposes	8.0	1.76	0.62	1.96	3.19
	Writing - Conventions of Standard English	8.0	2.02	0.58	2.26	3.46
	Writing - Research	8.0	1.49	0.69	1.57	2.81
	Listening	8.4	1.50	0.69	1.60	2.86

Table 74. Score Reliability of Mathematics Total Full Test and Reporting Categories

Level	Reporting Category	Mean # Items	Mean SEM	Reliability	RMSE	SD Theta
3	Total	32	0.26	0.82	0.32	0.75
	Algebra	8	0.58	0.71	0.59	1.09
	Number & Quantity	8	0.73	0.68	0.71	1.27
	Measurement & Data	8	0.75	0.63	0.81	1.34
	Geometry	8	0.87	0.56	0.94	1.42
4	Total	32	0.28	0.85	0.35	0.89
	Algebra	8	0.66	0.71	0.68	1.27
	Number & Quantity	8	0.76	0.72	0.73	1.37
	Measurement & Data	8	0.77	0.69	0.84	1.49
	Geometry	8	0.88	0.59	0.97	1.51
5	Total	32	0.31	0.81	0.39	0.91
	Algebra	8	0.74	0.70	0.75	1.38
	Number & Quantity	8	0.96	0.60	1.05	1.66

Level	Reporting Category	Mean # Items	Mean SEM	Reliability	RMSE	SD Theta
	Measurement & Data	8	0.84	0.64	0.93	1.54
	Geometry	8	0.95	0.60	1.02	1.61
6	Total	32	0.32	0.83	0.41	0.99
	Algebra	8	0.79	0.68	0.85	1.49
	Number & Quantity	8	0.83	0.71	0.84	1.57
	Measurement & Data	8	0.94	0.57	1.05	1.60
	Geometry	8	1.02	0.60	1.11	1.76
7	Total	32	0.37	0.83	0.50	1.20
	Algebra	8	1.00	0.65	1.12	1.90
	Number & Quantity	8	0.97	0.70	0.98	1.78
	Measurement & Data	8	1.02	0.66	1.10	1.89
	Geometry	8	1.17	0.59	1.29	2.01
8	Total	32	0.43	0.85	0.53	1.35
	Algebra	8	1.15	0.66	1.22	2.10
	Number & Quantity	8	1.14	0.70	1.10	2.02
	Measurement & Data	8	1.16	0.67	1.23	2.16
	Geometry	8	1.25	0.63	1.31	2.14

Table 75. Score Reliability of ELA Reading and Writing Testlet

Reporting Category	Grade	Mean # Items	Mean SEM	Reliability	RMSE	SD Theta
Total	3	51.2	0.36	0.86	0.45	1.17
	4	51.3	0.35	0.87	0.43	1.21
	5	51.4	0.34	0.88	0.43	1.25
	6	51.5	0.35	0.88	0.44	1.28
	7	51.5	0.43	0.87	0.55	1.55
	8	51.4	0.46	0.88	0.58	1.67
Reading: Key Ideas and Details	3	9.9	1.07	0.70	1.14	2.08
	4	9.8	0.87	0.74	0.92	1.79
	5	9.8	0.96	0.72	1.07	2.03
	6	9.7	1.04	0.69	1.20	2.16
	7	9.7	1.37	0.68	1.55	2.75
	8	9.7	1.41	0.69	1.59	2.86

Reporting Category	Grade	Mean # Items	Mean SEM	Reliability	RMSE	SD Theta
Reading: Craft Structure/Integration of Knowledge and Ideas	3	9.3	1.21	0.66	1.26	2.16
	4	9.5	1.20	0.66	1.27	2.18
	5	9.6	1.09	0.69	1.23	2.20
	6	9.7	1.14	0.67	1.29	2.24
	7	9.8	1.36	0.67	1.55	2.69
	8	9.7	1.30	0.70	1.45	2.64
Reading: Vocabulary Acquisition and Use	3	8.0	1.09	0.71	1.05	1.95
	4	8.0	1.11	0.73	1.02	1.95
	5	8.0	1.16	0.74	1.10	2.16
	6	8.0	1.05	0.75	1.02	2.05
	7	8.0	1.32	0.78	1.27	2.74
	8	8.0	1.63	0.76	1.46	2.98
Writing - Text Types and Purposes	3	8.0	1.25	0.62	1.32	2.14
	4	8.0	1.27	0.62	1.36	2.19
	5	8.0	1.18	0.63	1.26	2.07
	6	8.0	1.29	0.64	1.41	2.36
	7	8.0	1.64	0.61	1.73	2.77
	8	8.0	1.83	0.57	2.08	3.19
Writing - Conventions of Standard English	3	8.0	1.33	0.64	1.40	2.34
	4	8.0	1.35	0.66	1.42	2.44
	5	8.0	1.22	0.69	1.29	2.30
	6	8.0	1.32	0.64	1.44	2.38
	7	8.0	1.78	0.58	1.97	3.05
	8	8.0	2.06	0.56	2.28	3.42
Writing - Research	3	8.0	1.34	0.60	1.37	2.17
	4	8.0	1.27	0.67	1.26	2.20
	5	8.0	1.28	0.65	1.38	2.32
	6	8.0	1.44	0.57	1.60	2.45
	7	8.0	1.45	0.65	1.51	2.55
	8	8.0	1.49	0.70	1.50	2.75

Table 76. Score Reliability of ELA Reading Testlet

Reporting Category	Grade	Mean # Items	Mean SEM	Reliability	RMSE	SD Theta
Total	3	27.3	0.50	0.67	0.77	1.34
	4	27.2	0.47	0.67	0.77	1.34
	5	27.2	0.49	0.71	0.80	1.47
	6	27.0	0.50	0.72	0.78	1.45
	7	27.1	0.63	0.69	1.05	1.88
	8	27.1	0.65	0.72	1.02	1.91
Reading: Key Ideas and Details	3	9.8	1.03	0.71	1.08	2.01
	4	9.8	0.93	0.71	1.01	1.87
	5	9.8	0.99	0.72	1.10	2.08
	6	9.4	1.16	0.65	1.40	2.38
	7	9.5	1.40	0.67	1.63	2.81
	8	9.7	1.45	0.68	1.65	2.93
Reading: Craft Structure/Integration of Knowledge and Ideas	3	9.4	1.21	0.65	1.25	2.10
	4	9.4	1.21	0.66	1.33	2.27
	5	9.4	1.14	0.69	1.23	2.20
	6	9.6	1.13	0.68	1.24	2.21
	7	9.6	1.42	0.64	1.63	2.72
	8	9.4	1.41	0.68	1.54	2.72
Reading: Vocabulary Acquisition and Use	3	8.0	1.07	0.72	1.02	1.92
	4	8.0	1.05	0.73	1.00	1.92
	5	8.0	1.18	0.73	1.15	2.20
	6	8.0	1.09	0.74	1.06	2.06
	7	8.0	1.39	0.76	1.36	2.80
	8	8.0	1.70	0.75	1.57	3.11

Table 77. Score Reliability of ELA Writing – Text Types and Purposes Testlet

Reporting Category	Level	Mean # Items	Mean SEM	Reliability	RMSE	SD Theta
Writing – Text Types and Purposes	3	10.0	0.94	0.70	1.16	2.11
	4	10.0	0.87	0.75	1.04	2.06
	5	10.0	0.86	0.73	1.08	2.08
	6	10.0	0.87	0.75	1.08	2.16
	7	10.0	1.05	0.72	1.39	2.62
	8	10.0	1.19	0.73	1.51	2.91

Table 78. Score Reliability of ELA Writing – Conventions of Standard English Testlet

Reporting Category	Level	Mean # Items	Mean SEM	Reliability	RMSE	SD Theta
Writing – Conventions of Standard English	3	10.0	0.91	0.75	1.10	2.18
	4	10.0	0.90	0.78	1.07	2.26
	5	10.0	0.88	0.76	1.07	2.18
	6	10.0	0.91	0.73	1.15	2.21
	7	10.0	1.08	0.72	1.42	2.68
	8	10.0	1.31	0.70	1.74	3.20

Table 79. Score Reliability of ELA Writing – Research Testlet

Reporting Category	Level	Mean # Items	Mean SEM	Reliability	RMSE	SD Theta
Writing – Research	3	10.0	0.98	0.71	1.09	2.01
	4	10.0	0.88	0.78	0.97	2.06
	5	10.0	0.82	0.76	1.02	2.10
	6	10.0	0.93	0.70	1.23	2.25
	7	10.0	0.96	0.77	1.16	2.40
	8	10.0	1.10	0.76	1.33	2.72

Table 80. Score Reliability of ELA Listening Testlet

Reporting Category	Level	Mean # Items	Mean SEM	Reliability	RMSE	SD Theta
Listening	3	9.5	1.03	0.71	1.10	2.05
	4	8.8	0.98	0.73	1.02	1.97
	5	8.7	1.05	0.69	1.18	2.11
	6	8.6	1.14	0.72	1.20	2.24
	7	8.9	1.36	0.69	1.50	2.72
	8	9.3	1.57	0.67	1.70	2.98

Table 81. Score Reliability of Mathematics Testlets

Reporting Category	Level	Mean # Items	Mean SEM	Reliability	RMSE	SD Theta
Algebra	3	10.0	0.50	0.77	0.52	1.07
	4	10.0	0.51	0.78	0.56	1.17
	5	10.0	0.59	0.76	0.66	1.36
	6	10.0	0.64	0.74	0.74	1.45
	7	10.0	0.77	0.68	1.01	1.80
	8	10.0	0.95	0.70	1.14	2.09
Number & Quantity	3	10.0	0.61	0.70	0.71	1.29
	4	10.0	0.59	0.75	0.68	1.37
	5	10.0	0.68	0.67	0.90	1.56
	6	10.0	0.69	0.72	0.87	1.64
	7	10.0	0.78	0.74	0.95	1.85
	8	10.0	0.90	0.73	1.07	2.06
Measurement & Data	3	10.0	0.61	0.67	0.75	1.31
	4	10.0	0.57	0.76	0.68	1.39
	5	10.0	0.64	0.71	0.76	1.42
	6	10.0	0.72	0.67	0.86	1.50
	7	10.0	0.82	0.71	0.99	1.83
	8	10.0	0.96	0.69	1.20	2.15
Geometry	3	10.0	0.76	0.56	0.92	1.39
	4	10.0	0.72	0.63	0.89	1.46
	5	10.0	0.77	0.62	0.99	1.61
	6	10.0	0.88	0.55	1.22	1.82
	7	10.0	0.99	0.59	1.32	2.06
	8	10.0	1.03	0.65	1.29	2.17

Chapter 6

STANDARD SETTING

When students take a DRC BEACON test, they receive scale scores based on their performances. The scale score represents a quantitative, defensible measure of each student's level of knowledge and skills at the time of the test. However, the scale score alone does not indicate the knowledge and skills that the student likely has, nor does it indicate the skills a student must learn to progress. Most educators need support to put students' scale scores in context.

To help educators make sense of their students' test results, DRC BEACON reports a system of interrelated *performance levels* and *performance bands* that indicate the types of knowledge and skills that students likely have based on their test performances. These are key elements of the DRC BEACON *performance standards*, which are described in this section.

This section describes the process that was used to establish cut scores for the DRC BEACON ELA and mathematics tests; develop performance level descriptors (PLDs) for the tests; create nine *performance bands* for each reporting category from the test-level cut scores; and develop actionable, empirically based reporting details for each performance band. This section also offers guidance on the types of inferences that can be drawn from the DRC BEACON cut scores, performance levels, and performance bands.

Main Performance Levels and Cut Scores

At the test level, DRC BEACON reports a performance level for each student's performance: *Support Needed, Near Target,* or *Prepared.* These performance levels let educators quickly understand whether a student's current level of knowledge and skill is near the level needed for success in the next grade or course. Table 82 shows brief descriptions of these three performance levels for the tests.

Table 82. Description of the Three Main DRC BEACON Performance Levels

Performance Level	Description of Each Main DRC BEACON Performance Level
Support Needed	Students need support to gain the required skills for success in the next grade or course.
Near Target	Students are likely at (or near) the level of skill needed for success in the next grade or course.
Prepared	Students are likely prepared for success in the next grade or course.

The descriptions shown in the table are not specific to any particular test (e.g., grade 3 reading, grade 8 ELA). Instead, these *policy descriptors* are designed to be applied to any test.

For each test, the performance levels are defined by two *cut scores* on the relevant DRC BEACON scale. These cut scores were based on recommendations received from a nationwide committee of educators at the DRC BEACON *standard setting* study held in September 2018. This standard setting is summarized in the next subsection. Details about how the cut scores were adapted for DRC BEACON are presented in the following subsection.

It is important to note that the cut scores implemented for DRC BEACON are similar to—but not identical to—the cut scores used for the previous version of the test. These differences are also described in the following subsection.

2018 Standard Setting Study

On September 15–16, 2018, Data Recognition Corporation (DRC) conducted a standard setting study, in a workshop format, for the first generation of DRC BEACON tests. The standard setting engaged a diverse committee of educators from across the United States as they reviewed the content-based expectations for students "on track" to demonstrate college- and career-readiness by the end of high school, considered *benchmarks* based on analyses of the test data, and recommended cut scores for each test.

At the standard setting, participating educators considered a single cut score per test; during the workshop, this cut score was known as *On Track.* Accordingly, participants considered two performance levels: *On Track* and *Not On Track*. After the standard setting, DRC used this cut score to create three performance levels for DRC BEACON. This process is described in the next subsection.

Standard Setting Methodology and Rationale

A modification of the Bookmark Standard Setting Procedure (Lewis et al., 1996; Lewis et al., 2012) was implemented to establish cut scores for the tests. This method has been used on large-scale educational assessments across the nation (Karantonis & Sireci, 2006), including DRC's TABE and TASC tests.

As an item-mapping process, Bookmark is particularly useful for large-scale assessments such as DRC Beacon that include both multiple-choice and constructed-response items. Because Bookmark allows these different item types to be ordered together in *ordered item booklets*, and because of its history of use across the nation, DRC selected Bookmark for the standard setting.

Standard Setting Committee

Educators from across the nation participated in the standard setting. DRC recruited participants from the community of DRC BEACON users, as well as users of other DRC shelf and state assessments. DRC took special care to invite workshop participants who met the following criteria:

a) They were well qualified (e.g., had experience teaching in their associated content area).
b) They were diverse in terms of demographic characteristics (e.g., gender, ethnicity).
c) They were diverse in terms of geographic location.
d) They had knowledge of the tested content and population.

The standard setting committee comprised 16 educators. Of these, eight were assigned to the ELA committee and eight to the mathematics committee. Each group focused on all six tests in the relevant

content area. This structure was used to promote consistency of process and recommendations across grades.

The eight participants in the ELA committee were educators from Alabama, Arizona, California, Illinois, Maryland, and Mississippi. The eight members of the mathematics committee hailed from California, Iowa, Michigan, Minnesota, and Texas.

Participants were asked to describe their personal and professional backgrounds on the workshop evaluation. Of the 16 participants, 11 were female and five were male; 10 were caucasian four were Black, one was Asian, and one was Hispanic. Three participants were classroom teachers, three were non-teacher educators, four worked in higher education, and six were retired or held other positions.

Two facilitators from DRC helped guide each group through the standard setting process. These facilitators were members of the workshop staff and did not contribute to the recommendations.

Standard Setting Materials

Participants studied four key pieces of information at the workshop: performance level descriptors (PLDs), ordered item booklets (OIBs), item maps, and benchmarked cut scores.

Performance Level Descriptors (PLDs)

Using language from the DRC BEACON content standards, PLDs summarize the knowledge and skills expected of *On Track* students for each test. The PLDs were developed by an experienced team of content area experts and measurement staff at DRC.

To create the PLDs, DRC determined the knowledge and skills that students should have in order to be considered proficient in relation to the DRC BEACON college- and career-ready standards. DRC first made a determination as to the necessary content-based characteristics of students in the *On Track* performance level of DRC BEACON. Then DRC categorized these content-based characteristics for *On Track* students to create a clear, easily understood definition of *On Track* performance for each test. Finally, DRC made sure that the descriptions encompassed the performance continuum for *On Track* students, including knowledge and skills held by students at the threshold of the *On Track* level, in the heart of the level, and beyond. Created in this way, the definitions comprise *range PLDs*, detailing the knowledge, skills, and abilities expected of *On Track* students.

At the DRC BEACON standard setting, participants used these range PLDs to create informal *threshold PLDs*. To do this, participants considered the content-based expectations of students at the threshold (or point-of-entry) of the *On Track* performance level.

Ordered Item Booklets (OIBs)

An OIB was prepared for each test. Each OIB comprised items from the DRC BEACON test pool, all ordered in terms of difficulty. Item difficulty was calculated using data from students' performance on the tests using data collected in early and mid-2018. Easier items appeared earlier in the OIB, and harder items appeared later. Items ascended in terms of difficulty throughout the OIB. Multiple-choice (MC) and constructed-response (CR) items were ordered together in the OIB.

To order the items, a response probability criterion of 0.50 (RP50) was applied. When this criterion is applied, the RP50-adjusted scale location for an item is defined as the scale value associated with a 50% chance of answering the item correctly. DRC selected RP50 by comparing the RP-adjusted difficulty of the test items with the observed distributions of students' scores and observing that there was good overlap when RP50 was applied.

Each OIB comprised 10 items, purposefully selected to be near the benchmarked cut scores. Of the 10 items, 4 items were selected to be within a band of ±0.25 conditional standard error of measurement (CSEM) of the benchmarked cut score, 4 more items were within a band of ±0.50 CSEM, and 2 more items were within a band of ±1.0 CSEM. Half the items had RP50-adjusted scale locations (difficulty estimates) that were below the benchmarked cut score, and half had locations at or above the benchmarked cut score.

The OIBs were structured this way because participants were asked to react to the benchmarked cut scores and needed to examine items near the benchmarked cut scores to do so, they needed to examine items above and below the likely cut scores in terms of difficulty, and they needed to see items from different grades during the standard setting workshop to see the breadth of knowledge and skills measured by items across tests.

Item Maps

The item maps presented information for the items in the OIBs. The item maps showed each item's rank-order difficulty, RP50-adjusted scale location, scoring key, and the aligned content standard.

Benchmarked Cut Scores

To give participants a starting point for their judgments—and to allow participants to gauge the reasonableness of their recommendations—DRC presented the panelists with benchmarked cut scores. These benchmarked cut scores were presented in the form of pages in the OIBs. The use of benchmarks at standard settings is a well-documented way of providing policy-based and contextual information to standard setting participants (e.g., Phillips, 2012).

A benchmarked cut score for each test was calculated on the DRC BEACON scale using equipercentile methods. Using data from states that administered both DRC BEACON and the tests from a major multistate testing consortium, DRC used the DRC-leased items to find the points on the DRC BEACON test scale best aligned to the states' *Proficient* cut scores. For each test, the benchmarked cut scores were defined as the simple averages of the scale locations of these points. Calculated in this way, each benchmarked cut score referenced various states' *Proficient* cut scores in a statistically derived way. Given that DRC BEACON is frequently used to gauge whether students are on track to demonstrate proficiency at the end of the school year, DRC believed the cut scores for the tests would be similar to—but not necessarily exactly equal to—the benchmarked cut scores.

Participants at the standard setting were asked to discuss the content-based expectations for *On Track* students, to consider the benchmarked cut scores, and to make cut score recommendations for DRC BEACON. Participants were told how the benchmarked cut score for each test was calculated and that DRC believed the test data indicated that the final DRC BEACON cut score would likely be near the

benchmarked cut score. Participants were also told that the statistical calculations used to create the benchmarked cut scores were not enough: DRC needed participants to consider the content-based expectations for students in the *On Track* performance level, and participants needed to indicate whether the benchmarked cut scores reflected these expectations or whether different cut scores should be adopted.

Workshop Procedure

The DRC BEACON standard setting was conducted online in a two-day workshop. Prior to the workshop, participants were sent hard copies of the performance level descriptors (PLDs) and workshop materials.

In an opening session, DRC described the development of DRC BEACON, the process used to calculate the benchmarked cut scores, and the participants' main roles during the workshop. Specifically, participants were told their main roles were to discuss the content-based expectations associated with *On Track* students, consider the benchmarked cut scores, and recommend cut scores for DRC BEACON that were informed by these content-based expectations and benchmarked cut scores.

DRC then trained participants on the workshop methodology. Participants were shown training versions of the PLDs, OIB, and item map, and DRC described how they would be used during the workshop. Participants were told how they would study the OIBs, consider the benchmarked cut scores (in the form of pages in the OIB), discuss the content-based expectations for students in each performance level, and make cut score recommendations using the Bookmark Procedure. The committee then divided into one group for ELA and one group for mathematics. The steps the subgroups followed are listed below.

Grade 3. Participants began the standard setting process with grade 3. Afterwards, participants repeated the process for grades 4 through 8.

Threshold students. Participants studied the PLDs to consider threshold students, the hypothetical students with ability at the point-of-entry of the *On Track* level. Participants saw how the knowledge, skills, and abilities expected of *On Track* students were shown in the PLDs, and participants discussed their content-based expectations for threshold students.

Benchmarked cut scores. DRC introduced participants to the benchmarked cut scores. Participants were reminded how the benchmark was calculated. Participants were also reminded that their recommended cut scores would likely be near the benchmarked cut score, but that they would be asked to use their professional judgment to recommend a cut score that was consistent with the content-based expectations of *On Track* students.

Ordered item booklet (OIB). As a group, participants studied the items in the OIB. Starting with the easiest item, participants were asked (a) what each item measured, and (b) what made each successive item harder than the previous items. By asking these questions, participants gained a rich understanding of the knowledge and skills measured by the test items. DRC facilitated the conversation and took notes on an electronic item map: the item map was displayed alongside the items in the OIB.

Refresher training. During a supplemental training session, DRC reminded participants how bookmarks could represent cut scores in the OIB. Participants were reminded that their primary task was to make bookmark placements in the OIB that were consistent with the PLDs, with the tested content, and with their expectations for students. Participants were instructed to consider the benchmarked cut scores as they placed their bookmarks, and that they were free to recommend any bookmarks that were consistent with the knowledge and skills expected of the threshold students.

Participants were given a short quiz to gauge their understanding of the process. After answering questions, participants began the Round 1 of the Bookmark Procedure.

Round 1. Participants worked individually to place their bookmarks. To do so, participants were instructed to (a) consider the knowledge and skills measured by the items in the OIB the benchmarked cut score, and the content-based expectations of the threshold *On Track* student. Participants were instructed to start with a bookmark at the benchmarked cut score, consider the knowledge and skills measured by the items contained before that bookmark, and determine whether there was good correspondence between the content measured by the items before the bookmark and the expectations for the threshold *On Track* student. Participants would keep that bookmark if there was good correspondence or move the bookmark forward or backward in the OIB, one page at a time, until good correspondence was found.

DRC tabulated the Round 1 bookmarks and calculated each group's median cut score recommendations. DRC presented the participants with feedback based on their Round 1 bookmark placements, including the median bookmarks. Facilitators discussed the variability of participants' Round 1 bookmarks and how they compared with the benchmarked cut scores. DRC noted it was normal to have variability between participants when considering their Round 1 bookmarks, that consensus was not a goal, and that participants would have an opportunity to discuss their bookmarks with their colleagues.

Participants then shared their Round 1 bookmarks and the content-based rationales behind their bookmark placements. Participants were encouraged to refer to the OIB, item map, PLDs, benchmarked cut score, and threshold student expectations throughout this discussion.

Round 2. Following the discussion, participants again individually considered their bookmark placements. All participants made their bookmark placements individually and without discussion.

DRC tabulated participants' Round 2 bookmarks and calculated the median cut score recommendations. Participants' median cut score recommendation in Round 2 was taken as the group's recommended cut score. Table 83 shows the recommended cut scores from Round 2 of the standard setting as compared with the benchmarked cut scores. The differences between the recommended cut scores and the reference cut scores is given in terms of both scale scores and CSEM.

Table 83. Round 2 Cut Score Recommendations and Benchmarked Cut Scores for DRC BEACON

Content	Grade	Associated Benchmarks	Recommended Cut Scores	Diff. (SS Metric)	CSEM (Benchmark)	Diff. (CSEM Multiple)
ELA	3	415	412	-3	21.23	-0.14
	4	451	446	-5	20.75	-0.24
	5	471	471	0	19.83	0.00
	6	491	492	1	20.74	0.05
	7	501	504	3	23.41	0.13
	8	524	521	-3	26.4	-0.11
Mathematics	3	398	396	-2	36.02	-0.06
	4	459	455.5	-3.5	40.76	-0.09
	5	514	510	-4	44.52	-0.09
	6	530	530	0	43.95	0.00
	7	545	545	0	50.07	0.00
	8	579	579	0	58.31	0.00

Workshop Evaluation

Participants completed evaluations of the standard setting. Results from the evaluations can be used to gauge how fair and valid the participants perceived the standard setting process to be and whether participants supported their cut score recommendations.

Of the 16 participants, 15 completed evaluations at the end of the workshop. The evaluation results showed that participants understood the process and supported their recommendations. For example, participants were asked to indicate their level of agreement with the following statements. The percentage and number of participants who agreed or strongly agreed with each statement is shown in parentheses.

- The performance level descriptors (PLDs) were clear (100%, 15 out of 15).
- The PLDs communicate a reasonable profile of students' performance (100%, 15 out of 15).
- I am satisfied with my group's recommendations (93%, 14 out of 15).
- This process will produce valid performance standards for DRC BEACON (93%, 14 out of 15).

Acceptance of Educators' Recommendations by DRC

After the standard setting, DRC reviewed the recommendations from the educators who participated. DRC noted that participants' recommended cut scores were often different than the benchmarked cut scores and that these differences were usually small in terms of CSEM.

For grade 4 mathematics, the educator-recommended cut score was 455.5. To prevent the use of half-points on the test scale, DRC interpreted this recommendation as 455. DRC rounded the cut score down to the lower scale score to benefit students who score at that point on the test scale.

DRC noted that participants spent two days studying the DRC BEACON tests, talking about the expectations for *On Track* students, and creating cut score recommendations. Accordingly, DRC accepted the educator-recommended *On Track* cut scores for DRC BEACON.

Development of *Near Target* and *Prepared* Cut Scores

As previously stated, participants at the 2018 DRC BEACON standard setting considered a single cut score (then called *On Track*). Students who meet or exceed this level of performance are considered on track to be college- or career-ready by the end of high school.

To provide students and schools with more detailed feedback on students' performance on DRC BEACON, three performance levels were created using the *On Track* cut score at the heart of the system. These three performance levels, as introduced at the beginning of this section, are described here in greater detail:

- *Support Needed*. Students in this lowest performance level are still working to develop the knowledge and skills needed to be on track for college- and career-readiness, and they need support to gain the skills needed to succeed in the next grade or course. Their test performances are significantly below the *On Track* cut score.
- *Near Target*. Students in this middle performance level likely have the knowledge and skills needed to be considered on track for college- and career-readiness, and they likely have enough knowledge and skills to be successful in the next grade or course. Their test performances are just below, at, or just above the *On Track* cut score.
- *Prepared*. Students in this top performance level have the knowledge and skills associated with being on track for college- and career-readiness, and they are likely prepared for success in the next grade or course. Their test performances are significantly above the *On Track* cut score.

The *On Track* cut score always lies at the heart of the *Near Target* range. For each test, the *On Track* cut score is at the center of the scale range associated with the *Near Target* performance level. All students whose performance is classified as *Near Target* have earned scores near the *On Track* cut score.

The *Near Target* performance level was constructed this way to curtail potential overreliance on the *On Track* cut score. Because of measurement error associated with this (or any other) test instrument, any student's test score might be expected to be subtly different if they could be tested again with a parallel form of the test. With the *On Track* cut score at the center of the *Near Target* range, one can be reasonably confident that a student's level of knowledge and skills is truly below the *On Track* cut score

if the student is classified as *Support Needed* and that a student's skill level is truly above the *On Track* cut score if the student is classified as *Prepared.* For students in the *Near Target* range, their level of knowledge and skill is within a range commensurate with the *On Track* cut score.

In the previous generation of DRC BEACON tests, the bounds of the *Near Target* range (sometimes informally called the *green range*) were defined by the CSEM value associated with the *On Track* cut score for the test at hand. However, the CSEM values associated with the *On Track* cut scores for each test are not equal, and the scale-score differences between *On Track* cut scores for consecutive grades are also not equal. (This is to be expected: greater year-over-year growth in terms of knowledge and skills is typically expected of students in lower grades, and this is reflected in the cut scores.) This led to a pattern where the minimum scale score needed to be classified as *Near Target* was subtly lower in some grades than the previous grades, implying that less knowledge and skill were needed to be classified as *Near Target* in higher grades. Obviously, this is not the case, and a slightly different approach is used for the current DRC BEACON tests.

For current DRC BEACON tests, the *On Track* cut score still lies at the center of the *Near Target* range. To find the width of the *Near Target* range—the difference between the lowest and highest scale scores associated with the *Near Target* range for a given test—DRC found the *average* CSEM value associated with the *On Track* cut scores. The width of each *Near Target* range was defined as a fixed multiple of this average CSEM value. In ELA, the width of each *Near Target* range is defined as 161 scale score points, and in mathematics, the width of each *Near Target* range is defined as 151 scale score points. Constructed in this way, the minimum scores associated with the *Near Target* (and *Prepared*) performance levels rise monotonically, and the *Near Target* performance level is still centered on the educator-recommended *On Track* cut scores.

The scale ranges for DRC BEACON are shown in Table 84. These scale ranges were validated against newly developed PLDs for the DRC BEACON tests as described in the next subsections.

Table 84. DRC BEACON Scale Ranges for the Three Main Performance Levels

Content	Grade	Support Needed	Near Target	Prepared
ELA	3	160-331	332-492	493-800
	4	180-365	366-526	527-820
	5	200-390	391-551	552-840
	6	220-411	412-572	573-860
	7	240-423	424-584	585-880
	8	260-440	441-601	602-900

Content	Grade	Support Needed	Near Target	Prepared
Mathematics	3	160-320	321-471	472-800
	4	180-379	380-530	531-820
	5	200-434	435-585	586-840
	6	220-454	455-605	606-860
	7	240-469	470-620	621-880
	8	260-503	504-654	655-900

Development of PLDs for the Main Performance Levels

Although the DRC BEACON performance levels themselves signal whether a student's performance is considered "on track," educators typically need additional information to link these test results to real-world knowledge and skills. To this end, DRC has developed all-new performance level descriptors (PLDs) for DRC BEACON that reflect the scale ranges shown in Table 83, the items in the DRC BEACON pool, the DRC content learning progression, and the DRC BEACON content standards.

To create PLDs for the DRC BEACON tests, content experts from DRC used the DRC BEACON content standards and content learning progressions as their primary documents. These content experts began by examining the content standards to find the knowledge and skills expected of students in each grade and content area. Using the DRC BEACON content learning progressions, these content experts then summarized the knowledge and skills expected of students in the *Near Target* range (i.e., the knowledge and skills expected of students who are likely on track for college- and career-readiness and who would be ready for success in the next grade or course). Similar summaries were also developed for students in the *Support Needed* and *Prepared* performance levels.

Validating the DRC BEACON PLDS and Scale Ranges

The PLDs, performance levels, and scale ranges (i.e., cut scores) all form the *performance standards* for DRC BEACON. When examining DRC BEACON score reports, students and teachers should be reasonably confident that the elements of this system of performance standards all provide consistent messages about student performance. Accordingly, DRC validated the newly developed PLDs against the scale ranges. The purpose of this activity was to make sure the content claims made in the PLDs were consistent with the scale scores associated with each of the three main performance levels for the tests. The validation process is described below.

Item maps for each test. DRC began the process by mapping the DRC BEACON test items by difficulty. Similar to the approach taken during the 2018 standard setting, DRC mapped all the items associated with each test (e.g., grade 3 mathematics) by their scale locations (difficulty estimates). For two-point

items on the ELA test, each score point was mapped separately. A response probability criterion of 0.50 (RP50) was used, just as it was at the 2018 standard setting.

The scale ranges associated with each performance level were then shown on each item map. For example, the item map for grade 3 mathematics indicated which items had scale locations associated with the *Support Needed* level (i.e., 160–320), with the *Near Target* level (i.e., 321–471), and with the *Prepared* level (i.e., 472–800).

Comparing mapped items with PLDs. The scale locations associated with the item maps reflect empirical data collected on students' performance on the items. Accordingly, DRC content experts were asked to compare the DRC BEACON PLDs with the items "captured" in each performance level and to determine whether there was good correspondence between the PLDs and the items. To do so, the content experts considered the knowledge and skills needed by students to answer items in each performance level and then compared these skills with the PLDs. For example, the content experts considered the knowledge and skills needed to answer the items associated with the *Near Target* performance level in grade 6 ELA, and they compared these skills with the statements made on the grade 6 ELA PLDs for *Near Target*.

As needed, the DRC content experts refined the PLDs to reflect more clearly the types of knowledge and skills needed to answer the mapped items correctly. However, the content experts generally found good correspondence between the DRC BEACON scale ranges and PLDs, indicating there was good consistency between the descriptors and performance levels.

Developing Performance Bands for Reporting Categories

DRC recognizes that many educators would like to receive more granular information about students' strengths and areas of need. Specifically, educators have asked for students' performance to be described at the level of the *reporting category*. The 4–7 reporting categories established for each content area (e.g., Number & Quantity, Conventions of Standard English) can give teachers and stakeholders a more focused look at a student's knowledge and skills. Accordingly, DRC BEACON provides score reporting information, including performance standards, for each reporting category.

Educators have also indicated that they rely on the main performance levels for the tests, but they often wish more specific information about students' performance within each level was provided. For example, educators have asked for the specific knowledge and skills that students might need to develop to progress from the *Support Needed* level to the *Near Target* level. To provide this information, DRC has developed a new system of nine *performance bands* for each reporting category that align with the three main performance levels and provide actionable, data-driven information about student performance.

Creating the Nine Performance Bands

When a student takes a DRC BEACON test, that student receives a scale score for each reporting category. To help contextualize the student's performance, the scale score is accompanied by a performance band. As a special feature of the DRC BEACON performance standards, the performance band can help teachers and stakeholders better understand students' strengths and areas of need.

To create the nine performance bands for each reporting category, each of the three main DRC BEACON performance levels was divided into three bands. The resulting nine performance bands are used to report performance in each reporting category. The nine bands are summarized in Table 85.

As shown in Table 85, each trio of performance bands is associated with one of the main DRC BEACON performance levels. To create bands 1–3 and 7–9, the cut scores for the current grade and neighboring grades were used. Additional information about each trio follows the table.

Table 85. Description of Student Performance in Each of the Nine Performance Bands

Perf. Band	Level Alignment	Description of Student Performance in Each Performance Band
1	*Support Needed*	*Support Needed* for current grade and previous two grades.
2		*Support Needed* for current and previous grade, but *Near Target* for the grade before that.
3		*Support Needed* for the current grade, but *Near Target* for the previous two grades.
4	*Near Target*	The first quarter of the *Near Target* scale range for the current grade.
5		The middle half of the *Near Target* scale range for the current grade.
6		The final quarter of the *Near Target* scale range for the current grade.
7	*Prepared*	*Prepared* at the current grade and *Near Target* at the next grade.
8		*Prepared* at the current and next grade and *Near Target* at the grade after that.
9		*Prepared* at the current grade and next two grades.

Performance Bands 1–3: Support Needed in the Reporting Category

Students in these bands need help to learn the content needed for success in the next grade or course.

- **Band 1.** These students' performances are classified as *Support Needed* in the current grade. If cut scores from other grades were applied to these scores, they would also be classified as *Support Needed* in the previous two grades. (DRC BEACON defines cut scores for grades 3–8. When reporting scores for students tested at grades 3 and 4, the system extrapolates cut scores for grades 1 and 2 based on the other DRC BEACON cut scores.)
- **Band 2.** Performance at this level would be classified as *Support Needed* for the current grade and the previous grade. However, if the cut score from the twice-previous grade were applied, the performance would be described as *Near Target.*

- **Band 3.** Performance at this level would be described as *Support Needed* for the current grade. However, if cut scores from the two previous grades were applied, the performance would be described as *Near Target* in each.

Performance Bands 4–6: Near Target for the Reporting Category

In these three bands, students' performances are classified as *Near Target.* Students in these bands are at or near the target level of performance needed for success in the next grade or course.

- **Band 4.** In each grade, the entire *Near Target* scoring range spans 151–161 scale score points. The first quarter of this span (i.e., the first 38–40 scale score points) is associated with this band. Students in this band are near—but not yet at—the *On Track* cut score representing the target level of performance needed for success in the next grade or course.
- **Band 5.** The middle half of the *Near Target* scoring range (i.e., the middle 75–80 scale score points) is associated with this band. At the center of this band lies the implicit *On Track* cut score that denotes a level of knowledge and skills just sufficient to be associated with success in the next grade or course. Students at this level are at or near that level of knowledge and skill.
- **Band 6.** The final quarter of the *Near Target* scoring range (i.e., the highest 38–41 scale score points) is associated with this band. A student with performance in this band likely has enough knowledge and skill to be successful in the next grade or course.

Performance Bands 7–9: Prepared in the Reporting Category

Students in these bands have enough knowledge and skills to be successful in the next grade or course, and they are well prepared to meet the goals for future grades.

- **Band 7.** In this band, students' performances meet the *Prepared* target for the current grade. If cut scores from other grades were applied, the performances would be described as *Near Target* for the next grade.
- **Band 8.** Performance at this level would be classified as *Prepared* for the current and next grade, and as *Near Target* for the following grade.
- **Band 9.** A student in this band has performance that would be classified as *Prepared* when compared with the cut scores for the current grade and the next two grades. (When reporting scores for students tested at grades 7 and 8, DRC BEACON extrapolates cut scores for grades 9 and 10 based on the other DRC BEACON cut scores.)

As shown here, the performance bands can be used to gain a more nuanced view of students' performance in each reporting category. Although the performance bands themselves can be used to gain insight into a student's strengths and areas of need, educators often want to know exactly what kinds of skills the student likely has. More importantly, educators need to know the types of content the student needs to learn to progress. DRC BEACON meets this need by reporting information about the content associated with the nine performance bands.

Content Associated with the Performance Bands

When students and teachers view score reports through the *interactive reporting* feature, they are shown additional information associated with a student's performance in each reporting category, including the following:

- the performance band associated with the student's performance in that reporting category,
- a list of DRC BEACON content standards,
- contextual information associated with the circumstances under which a student can demonstrate reading skills, and
- the content standards and contextual information for the next performance bands.

This last item—the information for the next performance bands—shows the types of knowledge and skills a student needs to learn to increase performance in the reporting category.

This content-based information was developed through a data-centered, iterative process between DRC research scientists and content experts. The content-based claims made on these reports are grounded in the content learning progression used to create DRC BEACON. This content learning progression describes the key knowledge and skills that students learn across grades 3–8 and the typical sequence in which these skills are attained. The reports were also informed by the test data collected from thousands of test-takers nationwide. Details on how this information was derived is presented here.

Associating Content with Each Performance Band

As a computer-adaptive test (CAT), DRC BEACON has a large pool of test items. No two students are guaranteed to take the same items for a single reporting category, even if they receive similar scores. As a limitation, this means that the reports cannot present the exact skills measured by the items a student received, as this might reveal the contents of the confidential test items. However, it also means the pool of items for each reporting category can be used to make empirical claims about the types of knowledge and skills that are likely held by students in each performance band.

Vertical Scaling for DRC BEACON

DRC BEACON was created with a vertical scale. As such, students' scores in one grade can be directly compared with scores in neighboring grades. For example, if a student scores 500 in grade 3 mathematics, that score is associated with the same level of knowledge and skills as a student scoring 500 in grade 4 mathematics. Details about the precise methods used to create and maintain the vertical scale for DRC BEACON can be found in other sections of this report.

Similarly, item parameters are tied to the vertical scale. An item written for grade 5 can be administered to grade 5 students, but can also be administered to students in other grades to build a more precise picture of students' performance. This feature of the CAT is essential: off-grade items are frequently administered to students—especially to students with particularly low or high levels of performance—to calculate more precise test scores.

Item Maps for Each Reporting Category

Using the vertical scale, DRC built item maps for each reporting category. For each test item, a difficulty value was calculated based on DRC BEACON students' performance on the item. The RP50 criterion, as used at the 2018 standard setting, was used again.

The items from all grades for a given reporting category were then visualized on a single item map and ordered by difficulty. The nine performance bands for each grade were then projected onto each map, and the items captured in each performance band were identified. For example, an item map for all the items measuring the reporting category Geometry was created. To create the learning report for Geometry in grade 6 mathematics, the items with difficulty values associated with performance band 1 were found, followed by the items with difficulty values associated with performance band 2, and so on.

After this process was complete, items were associated with each combination of grade and performance band. These associated items all met the following criteria:

- They measured knowledge and skills associated with the reporting category.
- They had difficulty values consistent with the scale scores earned by students in the band.

DRC content experts then examined the items associated with each performance band. These content experts looked at the content standards to which each item was aligned. In consultation with the DRC BEACON content learning progression, they identified the content standards which were closely identified with each performance band. These standards denote the knowledge and skills that students in the performance band are mastering. The final list of standards includes a combination of standards that items in the item pool with difficulty values contained in the performance band measure directly, standards describing important precursor skills needed to respond successfully to items with difficulty values contained in higher performance bands, and standards denoted by the DRC BEACON content learning progressions as being essential to development across the performance continuum.

For each of the 594 combinations of grade, reporting category, and performance band, a list of content standards was compiled using this technique. An illustrative example of one of these combinations is presented in Figure 28. As shown in the figure, a list of skills is associated with a performance band in grade 5 mathematics. This list is not designed to be a canonical list of all the skills held by students in that performance band; instead, the list is designed to illustrate the general level of knowledge and skills typically held by students who score in that performance band for that reporting category. (Note that the full, interactive report includes a brief description of each content standard alongside the code. Fuller descriptions of the standards can be accessed through hyperlinks to the reports.)

Figure 28. Example of Content Standards Associated with a Performance Band

Grade 5 Mathematics: Algebra Reporting Category
Performance Band 5 (Scale Scores 573–547)

Students in this performance band have skills associated with the following standards: 4.OA.3, 4.OA.4, 4.OA.5, 5.OA.1, 5.OA.2, 5.OA.3, 6.EE.2, 6.EE.5, 6.EE.6

Brief descriptions of these standards provided on report, including links to expanded descriptions.

Compiled in this way, the lists of content standards associated with each performance band are based on actual test data—via the item maps—as informed by the DRC BEACON content learning progressions. Because of the nature of the content areas and content standards, additional contextual information is provided for reading.

Additional Context for Reading

For mathematics and writing, the item maps revealed an association between the performance bands and the content standards: items associated with certain content standards (especially those associated with higher grades) tended to be harder. Content experts noted that the content standards measured by those harder items required more knowledge and skills.

For reading, a slightly different pattern emerged. When mapped by difficulty, there was not always a clear association between the performance bands and the content standards. Notably, items associated with the various content standards were often associated with items of various difficulty levels. Part of this variation was due to the partial-credit nature of many reading items: students in different performance bands could demonstrate their partial knowledge of the skills associated with many content standards by earning partial credit on items associated with those standards. However, a clearer pattern emerged when content experts evaluated the complexity of the text in the passages/stimuli associated with items: students in higher performance bands could demonstrate the same skills as their peers in lower bands, but they could do so on more challenging texts.

Such a connection between text complexity and item difficulty was not a surprise. In fact, the DRC BEACON item pool was designed to incorporate passages of many different text complexity levels. To address this important facet of reading, DRC content experts created brief descriptions of the *context* under which students in various performance levels could demonstrate skills.

For example, the grade 5 reading standard 5.RI.1 specifies that students will “Quote accurately from a text when explaining what the text says explicitly and when drawing inferences from the text.” Clearly, students in many different performance bands may be able to quote from texts. However, a text’s complexity may affect the level of inferencing needed—necessary information may be explicit or implicit—and complex texts may challenge some students as they work to support their inferences with accurate evidence.

Figure 29 shows an example of the contextual information provided to help illustrate how students can demonstrate this skill at different levels. In this example, both performance bands 3 and 7 are

associated with reading standard 5.RI.1, as previously described. However, students in performance band 3 may be able to demonstrate this skill only on lower complexity texts related to familiar topics. In contrast, students in performance band 7 may be able to demonstrate this skill on grade-appropriate texts of moderate complexity. This contextual information helps educators better understand the circumstances under which students may demonstrate this reading skill.

Figure 19. Example of Contextual Information for Reading Performance Bands

Grade 5 Reading: Key Ideas & Details Reporting Category
Both of these performance bands are associated with standard 5.RI.1.

Performance Band 3 (Scale Scores 366–390)
The student demonstrates the ability to read and comprehend literary and informational texts at low complexity of familiar topics or themes.

Performance Band 7 (Scale Scores 552–572)
The student demonstrates the ability to read and comprehend literary and informational texts of moderate complexity.

Distinctions between Mathematics and ELA

As stated above, contextual information is shown only for selected reporting categories within ELA. This distinction between mathematics and ELA reveals an important difference.

For mathematics, the DRC BEACON standards (and associated standard codes) tend to be more granular and skills-based in nature. As students gain mathematical knowledge and skills, the items that measure their new skills tend to change in perceptible ways. This is reflected in the content codes shown in the performance band table: the content codes associated with higher performance bands tend to be associated with more complex skills. For example, content code 4.OA.1 is associated with students in bands 1 and 2: this standard involves interpreting a multiplication equation (e.g., 35=5*7) as a comparison. By band 3, 4.OA.1 is replaced by 4.OA.2, a standard which specifies using multiplicative comparison to solve word problems. Here, 4.OA.1 is a prerequisite skill to 4.OA.2, and this is mirrored in the performance band reporting: the easier skill is associated with lower bands, and the harder skill is associated with higher bands.

In contrast, some of the reporting categories within ELA do not have this quality. For example, content code 5.RL.3 is associated with a standard asking students to compare two or more characters or events by using details in a text. This skill is likely held, in some measure, by nearly all students who take the grade 5 reading test, and as such, this content code might be associated with nearly any band in the Key Ideas & Details reporting category. However, *the complexity of text* is a key distinction between students' performances in different performance bands. For example, a student in band 4 might be able to demonstrate this skill on texts of low-to-moderate complexity when the topic or theme is familiar, but students in band 6 might be able to demonstrate this skill on most low-to-moderate complexity texts, even if the topic or theme is new. This context is vital to understanding students' performances.

The list of content codes associated with each performance band was refined and amplified by DRC content experts. For students taking the tests in grades 3 and 4, the list of codes sometimes includes content from grades 1 and 2 for students in the lower performance bands. Analogously, students in higher performance bands in grades 7 and 8 will sometimes be shown content associated with high school. To select these content codes, DRC content experts relied on the DRC BEACON content learning progressions and the test data.

Content Associated with Higher Performance Bands

Although it is helpful to understand the types of skills that students in each performance band are currently working to master, many educators also want to know the skills that students need to grow. For this reason, the interactive reports for DRC BEACON also show the content standards (and for reading, contextual information) associated with the next two higher performance bands.

By examining the knowledge and skills associated with the next higher performance bands, educators can gain a better sense of the content students need master in order to increase their level of knowledge and skills.

Limitations on Inferences Made from Performance Bands

The DRC BEACON performance bands are designed to provide educators with actionable, data-based information about student performance. However, two limitations are suggested on the inferences that can be drawn from these performance bands. Notably, performance bands are not defined at the content standard level, and the performance bands are not equal-interval like the DRC BEACON test scale.

Performance Bands Are Not Defined at the Content Standard Level

As described in this section, performance bands are defined for each content area and reporting category. However, DRC BEACON does *not* currently support reporting at more granular levels than the reporting category, such as the content standard level.

The DRC BEACON tests are designed to be brief: when taking the test, only a handful of items are used to measure each reporting category. Accordingly, the reporting category is the most granular level of reporting in DRC BEACON.

From a theoretical perspective, DRC BEACON is an interim assessment designed to provide teachers and schools with defensible, quantitative information about students' progress over time. Although the system provides some reporting information on student performance (i.e., in each reporting category), DRC BEACON forms are not currently long enough to support reporting at the content standard level (e.g., 5.OA.3). Were scores reported at this level, several items would need to be administered for each standard, greatly expanding testing times. Accordingly, performance bands are not currently defined at the content standard level.

Performance Bands are Not Equal-Interval

The DRC BEACON test scales are *equal-interval*, and a difference on one range of the scale has the same theoretical interpretation as a difference of the same size elsewhere on the scale. For example, if a

grade 3 student were to gain 10 scale score points on the test between administrations (e.g., 400 to 410), this could be interpreted in approximately the same way as a grade 4 student who also gained 10 scale score points (e.g., 470 to 480). These differences can be compared because the scale has an equal-interval property: the value of one scale score point has the same theoretical meaning all along the test scale.

In contrast, the DRC BEACON performance bands do not have an equal-interval property. Each performance band has its own width (i.e., difference between minimum and maximum score), so gains in test performance quantified by performance bands cannot be directly compared. For example, imagine two students both increased by two performance bands in the Algebra reporting category of grade 4 mathematics. One student increased from band 3 (with a scale score of 375) to band 5 (with a scale score of 420), a difference of 45 points. Another student increased from band 6 (with a scale score of 495) to band 8 (with a scale score of 600), a difference of 105 points. Even through both these hypothetical students showed growth in Algebra, and even though both increased two performance bands, the amount of underlying growth along the test scale is not the same: the second student showed much more growth in Algebra than the first student.

For this reason, users of DRC BEACON are encouraged to consider growth along the test scale when comparing the scores of different students (or groups of students) and to consider the differences in scale scores when comparing growth in this way.

Chapter 7

SCORING AND REPORTING

This chapter of the *DRC BEACON Technical Report* provides a summary of the types of scores and reports available to users, as well as the major activities, purposes and uses associated with those scores.

Types of Scores

Scale Scores

A scale score indicating a student's total performance is determined for each content area on the DRC BEACON assessments. The overall scale score for a content area quantifies the performance being measured by that content area test. In other words, the scale score represents the student's level of performance, where higher scale scores indicate higher levels of performance on the test and lower scale scores indicate lower levels of performance.

The DRC BEACON scale scores summarize the level of student performance in ELA or mathematics. Classroom teachers may use these scores as evidence of student performance in these content areas. This technical report presents evidence that the scale scores are reliable indicators of student performance in ELA and mathematics and that classroom teachers and administrators may make valid inferences about student performance from these scale scores.

Reporting Category Scores

Reporting category scores are subscores based on important content categories within each subject area. The reporting category scores reflect primary structural elements in test blueprints and item development. Reporting category scores facilitate focus on more discrete categories of content.

The purpose of reporting subscores on DRC BEACON tests is to show the relationship between the overall performance being measured and the performance in each of the reporting categories for each student. Teachers may use these subscores as indicators of strengths and weaknesses of individual students, but they are best corroborated by other evidence, such as homework, class participation, or observation.

When a complete set of testlets within a subject area is completed within a thirty-day period, the reporting categories can be aggregated to the full content area scores. At the aggregate level, district and school administrators may use this information for activities such as designing curriculum and improvement planning.

Performance Levels and Performance Level Descriptors

Performance levels are reported at both the full test level (e.g., mathematics) and reporting category level (e.g., Algebra). The performance levels for subscores are divided into the following categories: *Support Needed* (i.e., orange range on the reports generated), *Near Target* (i.e., yellow range on the reports generated), and *Prepared* (i.e., green range on the reports generated). The *Near Target*, or yellow, range was set by taking the CSEMs of the subscores from a large sample of students, computing the average CSEM across the entire score distribution, multiplying the average CSEM by 1.25, and then adding or subtracting that value, on the DRC BEACON scale, from the On Track cut score between the *Support Needed* and *Prepared* performance levels. A student's performances in the ELA and mathematics tests are then reported as one of three levels of performance.

Performance level descriptors (PLDs) describe the knowledge and skills students in the *Support Needed*, *Near Target,* and *Prepared* performance levels should demonstrate with respect to the content standards. As described in the previous section of this report, the PLDs were developed by a team of

content area experts and measurement staff at DRC. DRC's team of content area experts and measurement staff have overseen the development of performance level descriptors for a number of large-scale assessment programs. The team has unique hands-on knowledge of the development of PLDs for large-scale assessments, including interim assessments like DRC BEACON. The process used to create the PLDs, including the ways in which empirical test data were used to validate the statements in the PLDs, is presented in Chapter 6 of this report. The reporting categories and measured standards associated with each of the DRC BEACON tests are presented in Table 1 and Table 2 of this manual.

Types of Reports

Effective and timely reporting is critical for all testing but especially for interim assessments. Stakeholders need to receive clear and actionable reports of student progress in a timely manner to adjust instructional strategies for student growth, whether those reports are for a full content area assessment or for the more focused testlets. DRC BEACON provides immediate (within an hour) reporting of student results and aggregated results (by school, class, or district). DRC BEACON reports include measurements of growth, and the interim scores can be used to predict a range of performance on the summative test at the end of the year and/or provide a comparison to national test scores as linking studies are conducted.

Individual student results will also be reflected in a printable PDF of the Individual Student Report (ISR) and within the DRC Interactive Reporting portal. Aggregated results will be updated hourly to reflect test submissions by various classrooms within a district. The ISR is a simple report interface to share with students and parents, designed with educator input. Notable features of the ISR include the following:

- It is presented in a single page.
- It can be presented in color or black and white.
- It presents a student's performance level in one of the following categories:
    - *Prepared*
    - *Near Target*
    - *Support Needed*
- It presents a student's performance in the reporting categories.
- Testlets roll up to a composite score as though a full content area was tested, if all testlets are completed within 30-day windows (allowing for flexible administration choices).

In the DRC Interactive Reporting portal, DRC BEACON results are delivered in a dynamic interactive reporting system that provides immediate access to individual results, roster reports, links to college- and career-ready standards, and reports about the strengths and weaknesses of individuals and groups of students. The interactive reporting system also offers the opportunity to disaggregate, categorize, and sort data as needed. The Standards and Learning Content Progressions reports provide indications of student strengths and weaknesses so that targeted improvement planning and support may take place in the classroom and in the home.

The DRC Interactive Reporting portal provides both individual and aggregated results based on client default settings. Graphical representations (vertical and horizontal bar graphs, line graphs, pie graphs, and scatterplots) display data in efficient ways. Each graphic also has the relevant data in a table format. Graphical representations can be printed as PDFs, and all tables can be exported as CSV, Excel, or PDF files. Users can dig deeper into all displays using a series of drop-down menus.

DRC BEACON offers the choice to administer full tests or shorter, more focused testlets, which can be combined for a full test score and are available in each of the reporting categories. For mathematics, these reporting categories are Algebra, Number & Quantity, Measurement & Data, and Geometry. The English language arts categories are Key Ideas & Details, Craft & Structure/Integration of Knowledge & Ideas, Vocabulary/Acquisition & Use, Informational Text, Literary Text, Text Types & Purposes, Research, and Conventions of Standard English.

Since students may take the entire mathematics assessment or English language arts assessment at one time, or the assessment can be taken as testlets, scores are provided accordingly—either as the score for the entire content area assessment or for each of the reporting categories. A student's interim scores can be used to predict a range of performance on a summative test based on the student's performance level.

The performance levels and descriptions are provided below.

Performance Level	Description of Each DRC BEACON Performance Level
Support Needed	Students support needed to gain the required skills for success in the next grade or course.
Near Target	Students are likely at (or near) the level of skill needed for success in the next grade or course.
Prepared	Students are likely prepared for success in the next grade or course.

Sample DRC BEACON Interactive Reports and Uses

As illustrated by the figure below, many interactive reports are available to teachers and administrators.

Classroom Reports	School Reports	District Reports
Highlight student group and individual results, performance over time, and learning progressions for areas on which to focus instruction	Summarize results across the school, enabling breakdowns by grade, teacher, and student demographics	Provide a district-level view of results to identify longitudinal trends within and across regions, schools, and student groups
Most useful to Teachers, Assessment Coordinators, supporting staff	Informative for Principals, Assessment Coordinators, school administrators	Intended for district leaders, administrators, and Dept. of Education staff
Examples: Individual Student Report; Student Dashboard; Group Performance	*Examples:* Class Comparison; School Summary; Disaggregate Summary	*Examples:* School Comparison; Assessment Insights

Each report is highly customizable, allowing users to make efficient use of their time.

To help users get started, DRC BEACON makes several reports available right away. These reports are described below.

Individual Results

Class Roster

- Provides a list of students with some identifying elements (e.g., Student ID, DOB) and corresponding test and score information in a tabular view
- Includes results from a single test session
- Allows users (e.g., teachers, school, district) to quickly sort student results in a specific sequence or filter down to a subset of students (e.g., certain performance levels, certain scale score ranges)
- Provides one-click access to ISRs for viewing, downloading, and printing
- Includes an option to view each content area (i.e., mathematics or ELA) separately or both combined, along with three predefined views for Overall Content, Subject Area, and Reporting Category

Longitudinal Roster

- Provides a list of students with some identifying elements (e.g., Student ID, DOB) and corresponding test and score information in a tabular view
- Includes results from multiple test sessions
- Allows users to quickly identify the change (i.e., growth or regression) in score and performance between different instructional and testing periods

- Allows users to sort student results in a specific sequence or filter down to a subset of students (e.g., certain performance levels, certain scale score ranges)
- Provides one-click access to ISRs for viewing, downloading, and printing
- Includes an option to view each content area (i.e., mathematics or ELA) separately or both combined and an option to specify a date range for relevant test sessions

Student Dashboard

- Acts as a dashboard of various measures and context related to student testing results
- Includes all test results for a single student
- Provides scale scores for all tests over time sorted by reporting category, and comparisons to mean scores for groups

Group Results

Group Performance

- Displays scores and identifying information for a specified student group
- Includes results for multiple test sessions
- Includes different charts to provide visual representations of student scores and the change in scores between test events over time
- Includes a Grade option that allows users to view the performance level cut points for different grade/content/category combinations
- Includes an option to view each content area (i.e., mathematics or ELA) separately along with a predefined view for each reporting category and a drill-down to a report that is a Group Learning Progression view

Comparison Report

- Provides summary information for a single test event
- Includes different charts to provide visual representations of the percentage of students whose scores fall within each performance level and includes comparisons to mean scores at different levels
- Includes predefined views for district, school, and class, and sections for each content area that was tested
- Provides users with underlying data from the charts and links to additional reports for each reporting category

Disaggregate Summary

- Provides summary information for a single test event
- Includes different charts to provide visual representations of the percentage of students within a demographic group whose scores fall within each performance level and includes comparisons to mean scores at different levels
- Includes sections for each content area tested
- Provides users with underlying data from the charts

Instructionally Focused Reports

Group Learning Content Progression

- Displays scores and identifying information for a specified student group
- Includes results for a single test session
- Includes different charts to provide visual representations of student scores within each reporting category and corresponding band from the learning progression
- Includes a Grade option that allows users to view the performance level cut points for different grade/content/category combinations
- Provides users with underlying data from the charts and a link to the Individual Learning Progression report

Individual Learning Progression

- Displays scores and identifying information for a single student
- Includes results for a single test session
- Features tabular views of student scores within each reporting category, the corresponding band from the learning progression, and the associated standards and descriptions
- Provides users with the band and standards that align with a student's scores (labeled "Tested Standards") along with the standards associated with the next band(s) in the progression (labeled "Standards for Growth")

Growth Projection

- Displays all test results for a single student
- Includes actual scores for a student, a school mean, and a district mean, along with a projected score for each (representing "growth")
- Includes different charts representing the same baseline information to satisfy user preferences (e.g., horizontal line, vertical column, grid, table)
- Includes sections for each content area tested

The reporting categories and measured standards associated with each of the DRC BEACON tests are presented in Table 1 and Table 2 of this manual.

Performance Bands

In the Individual Learning Progressions report, a student's performance for each reporting category is described using the nine performance bands. The report details the types of knowledge and skills the student has and the content the student needs to learn to progress. Specifically, the report includes:

- a list of content standards associated with the types of knowledge and skills that the student currently has;
- a performance band that describes the student's performance in relation to cut scores for the current grade (and, when needed, for nearby grades); and

- selected content standards associated with knowledge and skills that the student would need to obtain to progress into higher performance bands for that reporting category.

The performance bands are described in the previous section of this report. Note that performance bands are provided on the Individual Learning Progressions and Group Learning Content Progressions reports only at the level of the reporting category and above.

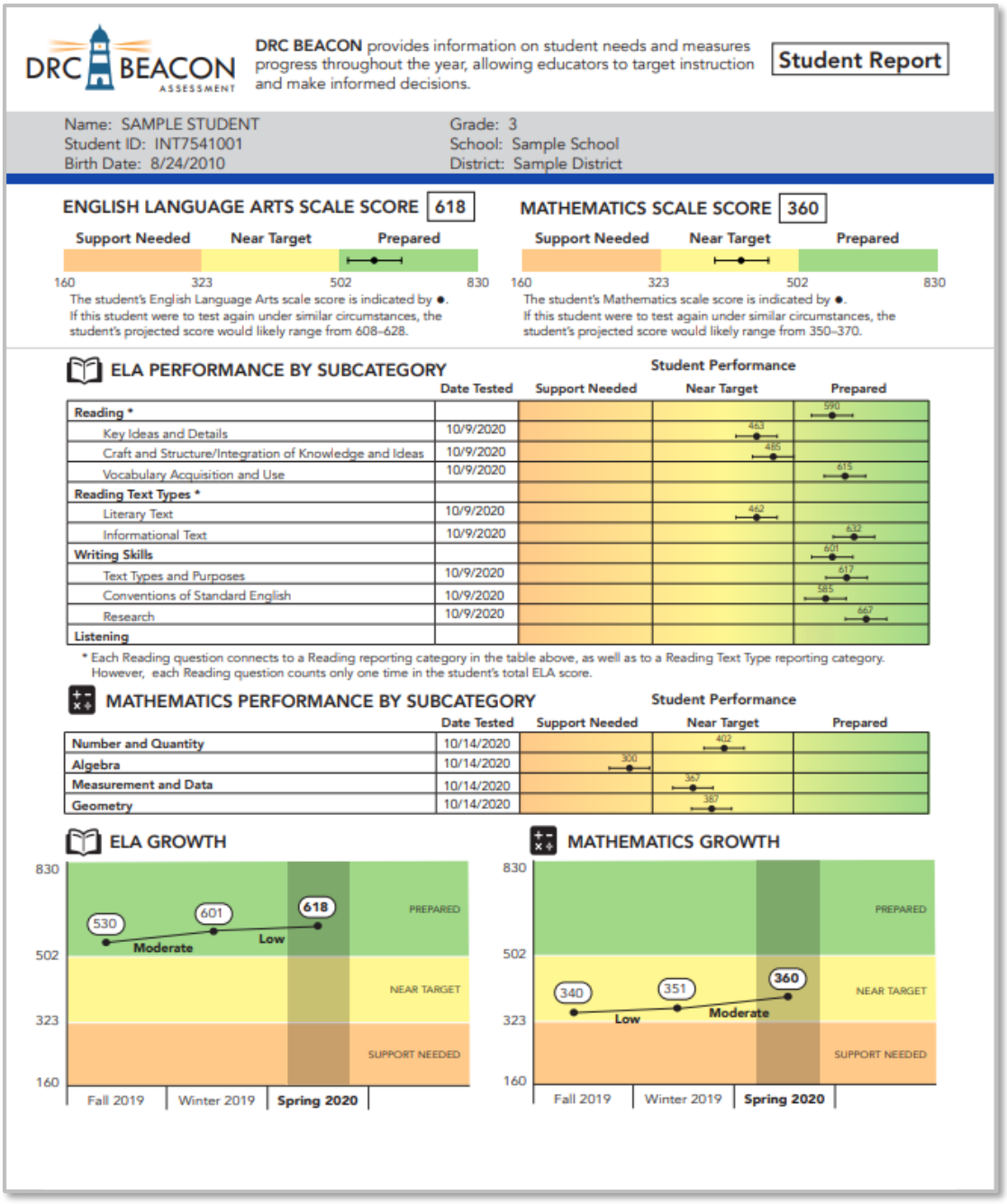
Individual Student Report

DRC BEACON supports parents and students who want to understand the progress that is being made. The Individual Student Report (ISR) can be generated for every student as often as is desired. The scores from testlets taken in a 30-day window will be rolled up to a total score.

The ISR can be a helpful tool for conferences with students as well as parents. The interactive dashboard can be viewed as a team, noting the variety of reports available for an individual student. The growth measure will allow for additional understanding and goal setting.

ISRs provide students and parents with information on overall performance and performance in each reporting category. Results populate the report based on the tests and testlets that have been administered for a composite score. The data on performance in reporting categories can be helpful as teachers seek information to target instruction.

A sample ISR is shown on the next page.



DRC BEACON ASSESSMENT

DRC BEACON provides information on student needs and measures progress throughout the year, allowing educators to target instruction and make informed decisions.

Student Report

Name: SAMPLE STUDENT
Student ID: INT7541001
Birth Date: 8/24/2010

Grade: 3
School: Sample School
District: Sample District

ENGLISH LANGUAGE ARTS SCALE SCORE 618

Support Needed | Near Target | Prepared

160 323 502 830

The student's English Language Arts scale score is indicated by ●.
If this student were to test again under similar circumstances, the student's projected score would likely range from 608–628.

MATHEMATICS SCALE SCORE 360

Support Needed | Near Target | Prepared

160 323 502 830

The student's Mathematics scale score is indicated by ●.
If this student were to test again under similar circumstances, the student's projected score would likely range from 350–370.

ELA PERFORMANCE BY SUBCATEGORY

	Date Tested	Student Performance: Support Needed / Near Target / Prepared
Reading *		
Key Ideas and Details	10/9/2020	483
Craft and Structure/Integration of Knowledge and Ideas	10/9/2020	485
Vocabulary Acquisition and Use	10/9/2020	615
Reading Text Types *		
Literary Text	10/9/2020	462
Informational Text	10/9/2020	632
Writing Skills		601
Text Types and Purposes	10/9/2020	617
Conventions of Standard English	10/9/2020	585
Research	10/9/2020	667
Listening		

* Each Reading question connects to a Reading reporting category in the table above, as well as to a Reading Text Type reporting category. However, each Reading question counts only one time in the student's total ELA score.

MATHEMATICS PERFORMANCE BY SUBCATEGORY

	Date Tested	Student Performance: Support Needed / Near Target / Prepared
Number and Quantity	10/14/2020	402
Algebra	10/14/2020	300
Measurement and Data	10/14/2020	367
Geometry	10/14/2020	387

ELA GROWTH

MATHEMATICS GROWTH

Using Educators' Input to Create the ISR

DRC used feedback from educators across the country to craft the look and functionality of the ISR. The ISR is designed to be easy to understand for teachers, parents, and students alike.

To create the ISRs, DRC first created six candidate formats for the reports. After internal review, four candidate reports were shared with a committee of educators. DRC and the reviewing educators had a common goal of creating ISRs that communicate meaningful, reliable information about student learning in a clear way. The educator review process is summarized below.

- An educator focus group was convened virtually.
- Participants represented five districts/grade levels, had ELA/mathematics classroom experience, and held a variety of leadership roles within their districts.
- DRC facilitated the focus group, and additional DRC staff members observed the session.

DRC gathered feedback from participants and facilitated discussion using a set of pre-developed guiding questions. The questions and follow-up discussions were designed to ensure that participants understood the reports and could freely provide input.

Focus Group Outline

A general outline of the focus group session is included below.

Activity
Welcome, introductions, ground rules, confidentiality
Overview of the purpose of the ISR
Explanation of the report review activity
Individual participant review of Page 1 of the report (same for all versions) with accompanying survey questions
Facilitated group discussion of consistent components
Individual participant review of Version 1
Facilitated group discussion of Version 1
Individual participant review of Version 2
Facilitated group discussion of Version 2
Individual participant review of Version 3
Facilitated group discussion of Version 3
Individual participant review of Version 4
Facilitated group discussion of Version 4
Discussion of any additional thoughts and remaining feedback
Thank you and close session

Guiding Questions for the Focus Group

The focus group concentrated on exploring educator, student, and parent/guardian needs regarding the score reports. The goal was to determine whether educators and parents/guardians would receive the information they needed in an easy-to-understand way. The session focused on both visual features (e.g., layout, format, flow, appearance) and content within the reports.

Questions for focus group participants included the following:

Initial overall reactions

- Do all the sections work together?
- How easy was it to move through the report, find the next section, and find supporting material to understand the score information?

Expected reactions by report users

- How would other educators, parents/guardians, or students react to the report?
- Any foreseen problems? Suggestions?

Report Text

- Overall, how easy is the report to read?
- Is the information relevant?
- Is the font type easy to read? Is the font large enough? Do the colors or background make the font difficult to read?

Wording

- Is the wording clear?
- What wording is too complex? Anything that could confuse users?
- Are there too many or too few words?

Interpretation of the data (scores)

- How difficult was it to learn how to read the report? (e.g., was it easy to determine what the scale score and growth measures meant?)
- Were the student's overall results and performance level clearly represented?

Other report elements

- Was the format (e.g., displays, color usage) consistent and helpful throughout the report?
- Was it clear where to go for help to interpret the report? Was enough information given on the report?
- Was there anything not included on the report that should be?

DRC facilitated a discussion about each candidate report with the educators. Afterward, the group compared and contrasted the four reports and shared insights.

Three follow-up sessions occurred with smaller subgroups of educators. Each follow-up group focused on a specific content area (i.e., ELA or math) or grade range (3–5 or 6–8). Insights that the educators provided that were particularly helpful included strong preferences for certain color combinations, a preference for a single page report, and insight into clarity of language.

Chapter 8

REFERENCES

Achieve.org. www.Achieve.org.

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

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Addison-Wesley.

Burket, G. R. (2002). *PARDUX* [Computer program]. Data Recognition Corporation.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Sage.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Sage.

Council of Chief State School Officers State Collaborative on Assessment and Student Standards: Technical Issues in Large-Scale Assessment (2003). *Quality Control Checklist for Item Development and Test Form Construction*. Washington, DC.

Council of Chief State School Officers (2013). *Navigating text complexity.* Navigating text complexity. http://navigatingtextcomplexity.kaulfussec.com/

Council of Chief State School Officers (2019). *Criteria for procuring and evaluating high-quality assessments*. Author.

EDL Core Vocabularies (1989). Steck-Vaughn.

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

Equity (2016). *Evaluating item quality in large-scale assessment: Phase 1 report of the study of state assessment systems*.

Herman, J., & Linn, R. (2015). *Evidence-centered design: A summary*. Curriculum Revision and Educational Standards for Teaching
http://cresst.org/

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

Jensen, A. R. (1980). *Bias in mental testing*. The Free Press.

Johnstone, C. J., Altman, J., & Thurlow, M. (2006). *A state guide to the development of universally designed assessments*. University of Minnesota National Center on Educational Outcomes.

Karantonis, A., & Sireci, S. G. (2006). The Bookmark standard setting method: A literature review. *Educational Measurement: Issues and Practice*, *25*, 4–12.

Karkee, T., Lewis, D. M., Hoskens, M., Yao, L., & Haug, C. (2003, April). *Separate versus concurrent calibration methods in vertical scaling* [Paper presentation]. The Annual Meeting of the National Council on Measurement in Education, Chicago, IL, United States.

Karkee, T., Wang, Z., Green, D. R., & Patz, R. J. (2006, April 10). *Vertical scaling of English language proficiency assessments using common examinees design: A comparison of three methods* [Paper presentation]. The National Council on Measurement in Education, San Francisco, CA, United States.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996). *Standard setting: A Bookmark approach* [Symposium]. The Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ, United States.

Lewis, D. M., Mitzel, H. C., Mercado, R. L., & Schulz, E. M. (2012). The bookmark standard setting procedure. In G. J. Cizek (Ed.*), Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 225–253). Routledge.

McClarty, K. L., Way, W. D., Porter, A. C., Beimers, J. N., & Miles, J. A. (2013). Evidence-based standard setting: Establishing a validity framework for cut scores. *Educational Researcher*, *42* (2), 78–88. https://doi.org/10.3102/0013189X12470855

Mogilner, A. (2006). *Children's writer's word book*. Writer's Digest Books.

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

National Assessment of Education Progress. (2019). *Mathematics framework*. https://www.nagb.gov/naep-frameworks/mathematics.html

National Assessment of Education Progress. (2019). *Reading framework*. https://www.nagb.gov/naep-frameworks/reading.html

National Center for Education Statistics (2010). *2002 educational longitudinal study: Second follow-up, 2006*.

Partnership for Assessment Readiness for College and Careers (PARCC).

Phillips, G. W. (2012). The benchmark method of standard setting. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 323–346). Routledge.

SMARTER Balanced and the ELA Council of Chief State School Officers SCASS.(2012) *English language arts content specifications.*

Thompson, S., Johnstone, C. J., & Thurlow, M. L. (2002) *Universal design applied to large scale assessments* (Synthesis Report 44). University of Minnesota, National Center on Educational Outcomes.

Webb, N. L. (1997). *Criteria for alignment of expectations and tests in mathematics and science education* (NISE Research Monograph No. 6). University of Wisconsin–Madison.

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states* (NISE Research Monograph No. 18). University of Wisconsin–Madison, National Institute for Science Education.

Webb, N. L. (2005, November, 19. *Depth-of-knowledge levels for four content areas* [Paper presentation]. The Florida Education Research Association, 50th Annual Meeting, Miami, FL, United States.

WestEd (2015). *Item specifications guidelines.* Council of Chief State School Officers.

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

