



# Evaluating the Content and Quality of Next Generation High School Assessments

## Final Report

**Prepared for:** Rockefeller Philanthropy Advisors  
6 West 48<sup>th</sup> Street, 10<sup>th</sup> Floor  
New York, New York 10036

**Authors:** Sheila R. Schultz  
Hillary R. Michaels  
Rebecca Norman Dvorak  
Caroline R.H. Wiley

**Prepared by:** Human Resources Research Organization  
66 Canal Center Plaza, Suite 700  
Alexandria, Virginia 22314

**Date:** February 11, 2016

# Evaluating the Content and Quality of Next Generation High School Assessments

## Table of Contents

Acknowledgements.....	iv
Executive Summary .....	v
Study Overview.....	v
Review Process .....	vii
Review Materials.....	viii
Reviewers .....	ix
Rating System.....	ix
Accessibility Review.....	x
Modifications from Prescribed Methodology .....	x
Summary of Results.....	x
Program Responses to Study.....	xviii
Conclusions .....	xviii
Chapter 1: Introduction .....	1
Background.....	1
Overview of Study.....	1
Parallel Study.....	2
Participating Assessment Programs .....	2
Organization of Report.....	4
Chapter 2: Overview of Test Content Evaluation Methodology .....	5
Study Criteria .....	5
Review Process .....	7
Review Materials.....	9
Reviewers .....	9
Rating System.....	10
Accessibility Review.....	11
Modifications from Prescribed Methodology .....	11
Chapter 3: Implementation of the Review Process.....	14
Evaluation Criteria.....	14
Review Materials.....	15
Reviewer Selection .....	18
Review Activities .....	18
Accessibility Review.....	21
Access to Exemplar/Sample Items .....	22
Scoring Procedures .....	23
Development of Summary Statements .....	24

## Table of Contents (Continued)

Chapter 4: Test Content and Depth Results .....	25
ELA/Literacy.....	25
Mathematics.....	32
Summary of Findings .....	36
Program Responses to Study .....	38
Chapter 5: Accessibility Results.....	39
ACT Aspire.....	39
MCAS.....	40
PARCC .....	41
Smarter Balanced .....	42
Accessibility Feature Comparison.....	43
Chapter 6: Study Challenges and Recommendations .....	46
General Challenges .....	46
Challenges with Specific Sub-criteria.....	47
Challenges with Accessibility Review .....	51
Concluding Commentary .....	51
Appendix A: CCSSO Criteria for High-Quality Assessments.....	A-1
Appendix B: ELA/Literacy Scoring Template .....	B-1
Appendix C: Mathematics Scoring Template.....	C-1
Appendix D: Accessibility Scoring Template.....	D-1
Appendix E: Metadata for Test Content Evaluation Methodology .....	E-1
Appendix F: Reviewer Biographies.....	F-1
Appendix G: ACT Aspire Criteria B and C Ratings and Summary Statements.....	G-1
Appendix H: MCAS Criteria B and C Ratings and Summary Statements.....	H-1
Appendix I: PARCC Criteria B and C Ratings and Summary Statements .....	I-1
Appendix J: Smarter Balanced Criteria B and C Ratings and Summary Statements .....	J-1
Appendix K: Testing Program Responses to HQAP High School Study .....	K-1
Appendix L: ACT Aspire Accessibility Summary Statements.....	L-1
Appendix M: MCAS Accessibility Summary Statements .....	M-1
Appendix N: PARCC Accessibility Summary Statements.....	N-1
Appendix O: Smarter Balanced Accessibility Summary Statements .....	O-1

## Table of Contents (Continued)

### List of Tables

Table ES1. Key Characteristics of the Four Assessments Included in this Evaluation Study .....	vi
Table ES2. Summary of Four Programs’ High School ELA/Literacy and Mathematics Ratings .....	xi
Table 1. Key Characteristics of the Four Assessments Included in the Evaluation Study.....	3
Table 2. Subset of CCSSO Criteria Implemented in Study of High School Assessments .....	6
Table 3. Illustration of Criterion C.1 with Sub-Criteria, Scoring Guidance, and Tentative Cut-Off.....	14
Table 4. Composite ELA/Literacy Content Ratings.....	25
Table 5. Rating for Close Reading (Criterion B.3) .....	26
Table 6. Rating for Writing (Criterion B.5).....	27
Table 7. Rating for Vocabulary and Language Skills (Criterion B.6).....	28
Table 8. Rating for Research and Inquiry (Criterion B.7).....	28
Table 9. Composite ELA/Literacy Depth Ratings.....	29
Table 10. Rating for Text Quality and Types (Criterion B.1) .....	30
Table 12. Rating for Cognitive Demand (Criterion B.4) .....	31
Table 13. Rating for High-Quality Items and a Variety of Item Types (Criterion B.9).....	32
Table 14. Composite Mathematics Content Ratings.....	32
Table 15. Rating for Focus (Criterion C.1) .....	33
Table 16. Rating for Concepts, Procedures, and Applications (Criterion C.2).....	33
Table 17. Composite Mathematics Depth Ratings.....	34
Table 18. Rating for Connecting Practice to Content (Criterion C.3) .....	34
Table 19. Rating for Cognitive Demand (Criterion C.4) .....	35
Table 20. Rating for High-Quality Items and a Variety of Item Types (Criterion C.5) .....	36
Table 21. Comparison of Select ELA/Literacy and Mathematics Accessibility Features across the Four Programs.....	44

### List of Figures

Figure ES1. Overview of rating process. ....	viii
Figure 1. Overview of rating process. ....	8

## Acknowledgements

This important work was possible from funding by the High Quality Assessment Project (HQAP), which supports state-based advocacy, communications, and policy work to help ensure successful transitions to new assessments that measure K–12 college- and career-readiness standards. HQAP’s work is funded by a coalition of national foundations, including the Bill & Melinda Gates Foundation, the Lumina Foundation, the Charles and Lynn Schusterman Family Foundation, the William and Flora Hewlett Foundation, and the Helmsley Trust.

We sincerely appreciate the cooperation and efforts of the testing programs that participated in the study—ACT Aspire, Massachusetts Comprehensive Assessment System, the Partnership for Assessment of Readiness for College and Careers, and the Smarter Balanced Assessment Consortium. In particular, we thank Elizabeth (Beth) Sullivan, Carrie Conaway, Francine Markowitz, Judy Hickman, and Nikki Elliott-Schuman. We also thank the many individuals who completed such thorough and careful reviews of the programs’ items and documentation.

## Executive Summary

In the new era created by the federal *Every Student Succeeds Act* signed into law in December 2015, states have the responsibility to ensure their educational systems produce students who are prepared for the worlds of higher education and work. Student assessments are the primary mechanism for gauging the success of state and local educational systems. There are a variety of assessments that can be used, including those developed by individual states, consortia of states, and commercial vendors. Each of these assessments will have particular strengths and weaknesses that make them more or less suitable for specific applications. The question then becomes, how can policy-makers obtain clear, thorough, and unbiased characterizations of particular student assessments that reflect the complexities of next generation testing goals, strategies, and formats?

### Study Overview

The National Center for the Improvement of Educational Assessment (NCIEA, hereafter referred to as the Center) developed an innovative evaluation methodology to address this complex question.<sup>1</sup> The methodology goes well beyond traditional studies that examine the alignment between discrete test items and learning objectives. It takes as its guiding framework elements of the *Criteria for Procuring and Evaluating High Quality Assessments*, which was developed by the Council of Chief State School Officers (CCSSO) and released in 2014. CCSSO developed its criteria to be applicable to any assessment that was intended to measure college- and career-ready content standards in mathematics and English language arts (ELA)/literacy, especially the Common Core State Standards (CCSS).

The Center's methodology translates the CCSSO criteria into specific rubrics and scoring procedures to facilitate both a credible and a practical evaluation of an assessment. To facilitate development of its methodology, the Center divided the CCSSO criteria into two parts: Test content and test characteristics. The test content evaluation procedures, which are the focus of the present study, highlight the extent to which an assessment (a) aligns to content standards, (b) is accessible to all students, and (c) is transparent in its test design.

The present study used the new methodology to evaluate the extent to which the high school ELA/literacy and mathematics summative assessments for four programs—ACT Aspire, the Massachusetts Comprehensive Assessment System (MCAS), Partnership for Assessment

---

<sup>1</sup> [http://www.nciea.org/publication\\_PDFs/Guide%20to%20Evaluating%20CCSSO%20Criteria%20Test%20Content%2001%2024%2016.pdf](http://www.nciea.org/publication_PDFs/Guide%20to%20Evaluating%20CCSSO%20Criteria%20Test%20Content%2001%2024%2016.pdf)

of Readiness for College and Careers (PARCC), and Smarter Balanced Assessment Consortium (Smarter Balanced)—match the CCSSO criteria relevant to test content. Key differences in the assessments for the four programs are outlined in Table ES1.

**Table ES1. Key Characteristics of the Four Assessments Included in this Evaluation Study**

Program	Subjects Reviewed	Mode	Grade Reviewed	Testing Time
ACT Aspire	Mathematics ELA/literacy - English - Reading - Writing	Online	Grade 10	3 hrs, 15 mins
2014 MCAS	Mathematics ELA/Literacy	Paper-Pencil	Grade 10	3 hrs, 30 mins
PARCC	Mathematics - Mathematics III - Algebra II ELA/Literacy	Online	Grade 11	7 hrs, 30 mins <sup>2</sup>
Smarter Balanced	Mathematics ELA/Literacy	Online, adaptive	Grade 11	5 hrs, 30 mins

<sup>2</sup>The 2015-2016 revisions will reduce this by an estimated one and one-half hours.

We addressed the following questions in the current study:

- Do the assessments place strong emphasis on the most important content of college and career readiness as called for by the Common Core State Standards and other college and career-readiness standards? (Content)
- Do the assessments require all students to demonstrate the range of thinking skills, including higher-order skills, called for by those standards? (Depth)
- Are the assessments accessible to all students, including English learners (ELs) and students with disabilities (SWDs)? (Accessibility)

The four assessment programs included in the study were intended to represent an array of options states might be considering for adoption. MCAS was included as it represents what has been considered “best in class” for individual state assessments up until this point.

<sup>2</sup> The 2015-2016 revisions will reduce this by an estimated one and one-half hours.

A parallel study was conducted by the Thomas B. Fordham Institute (hereafter referred to as Fordham), which implemented the Center’s methodology for grades 5 and 8 summative mathematics and ELA/literacy assessments. Taken together, HumRRO and Fordham were first to implement the Center’s evaluation methodology. HumRRO and Fordham conducted their studies separately; however, the two organizations communicated often about the evaluation methodology and collaborated on the steps to implement it.

### ***Review Process***

The methodology for evaluating assessments against the CCSSO test content criteria is a progressive process in which each step builds on the last step. To begin, individual subject matter experts review materials associated with the assessment program and judge the extent to which they match the applicable CCSSO sub-criteria. The ratings are successively aggregated culminating in two composites reflecting content and depth. In addition to ratings, reviewers at each stage of the process provide narrative comments that explain their ratings and highlight program strengths and weaknesses.

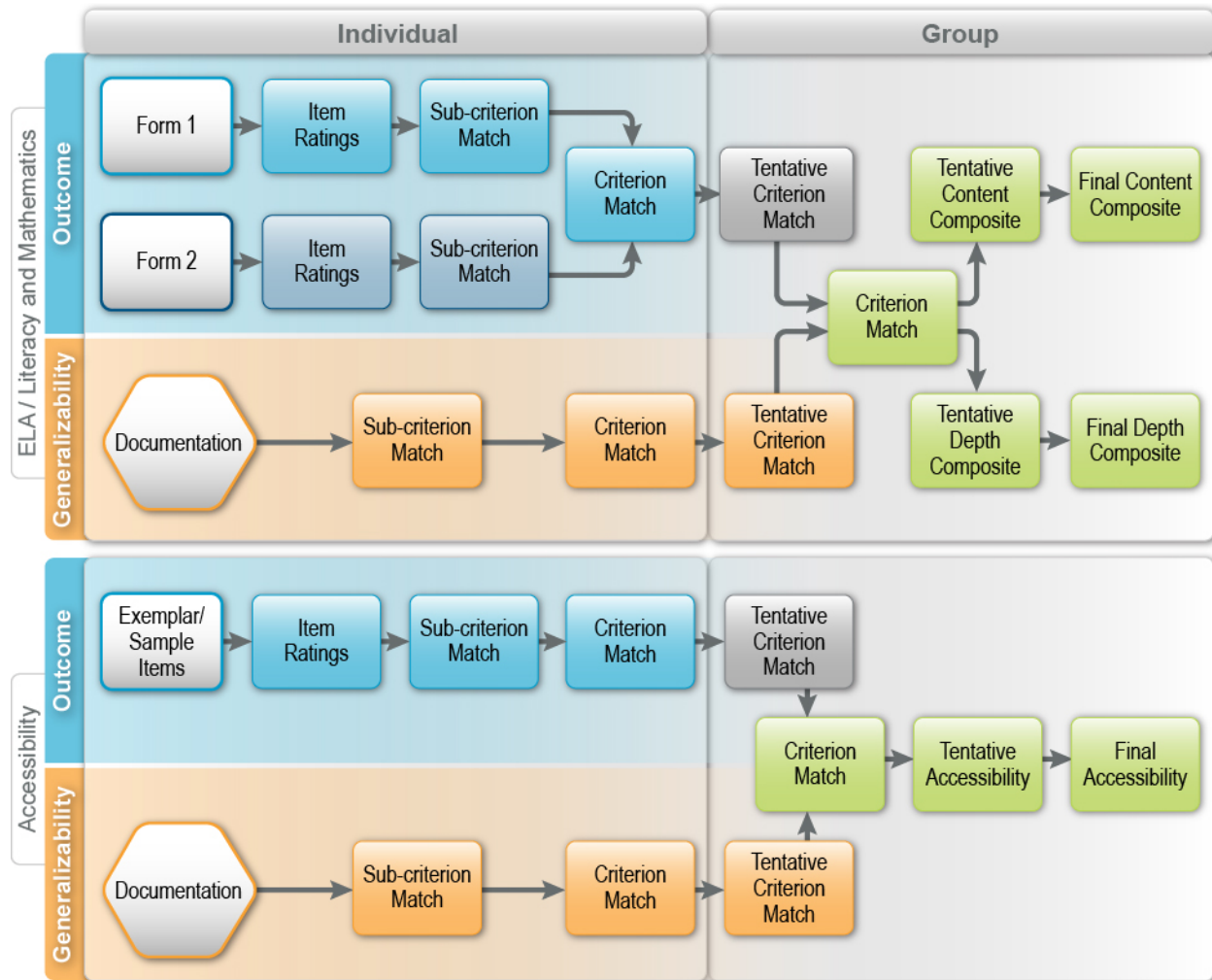
The methodology calls for the following four types of reviews:

- (1) Reviewing items and passages from two test forms to evaluate the extent to which operational assessment forms meet the CCSSO criteria (Outcome review).
- (2) Reviewing test documentation and specifications to evaluate the extent to which results are generalizable across forms of the assessment (Generalizability review).
- (3) Reviewing the extent to which the assessment program provides sufficient information to the public regarding assessment design and expectations (Transparency review).
- (4) Reviewing the extent to which the assessment program’s tests are fair to all students, including ELs and SWDs (Accessibility review).

The so-called Outcome and Generalizability reviews to evaluate test content and depth were conducted separately by HumRRO (for the high school assessments) and Fordham (for grades 5 and 8), and followed the Center’s methodology quite closely. The Transparency review was conducted jointly by HumRRO and Fordham, and involved review of documentation that proved to be so voluminous that a suitable evaluation could not be completed. The Accessibility review was conducted by HumRRO, with support from Fordham, and involved review of both Outcome



(exemplar test items) and Generalizability (documentation). This review also posed some challenges and yielded evaluations of each program but no summary ratings. Figure ES1 provides an illustration of the methodology. Further details about the methodology and how it was implemented in this study are provided in the main body of this report.



**Figure ES1. Overview of rating process.**

### Review Materials

The Center’s test content evaluation methodology requires two types of evidence to be examined to make judgments about the quality of an assessment program. The first type of evidence comes from examination of assessment items from operational test forms—this provides direct evidence of what students will have experienced and is referred to as Outcome evidence because the items and forms are the outcomes of an extensive design and development process. The second type of evidence comes from examination of documentation provided by the assessment program—this is known as Generalizability evidence because it helps determine

whether results from a limited review can likely be generalized across all test forms that a program might create. Operational forms and item and passage metadata (e.g., text complexity) provided the Outcome evidence while program documentation provided the Generalizability evidence.

### **Reviewers**

As indicated in Figure ES1, the Center’s test content evaluation methodology relies heavily on expert judgment; therefore, we paid careful attention to ensuring that highly qualified, unbiased individuals served as reviewers in the study. After reviewing their qualifications, we selected 20 individuals—content experts in ELA/literacy and mathematics, experienced classroom teachers, individuals with expertise in large-scale assessment, and experts in accommodating SWDs and ELs—to serve as Outcome reviewers (10 ELA/literacy reviewers and 10 mathematics reviewers). We also selected four individuals to serve as Generalizability reviewers; all of the individuals we selected as Generalizability reviewers also served as Outcome reviewers. The Generalizability review was conducted jointly with Fordham, so they also selected four individuals for this review. Finally, we selected nine different individuals to serve as Accessibility reviewers. Additional information about the reviewer selection process and the individuals who participated in our study is provided in Chapters 2 and 3 of this report.

### **Rating System**

The Center’s methodology uses a 3-point rating system (0, 1, 2) to reflect the match of the material and items reviewed to the CCSSO criteria. These “match scores” are accompanied by narrative comments provided by the reviewers that further characterize and explain their judgments. Each group-level evaluation (sub-criterion, criterion, and composite) is translated into one of the following labels to characterize the match to the CCSSO criteria:

- Excellent Match (match score = 2)
- Good Match
- Limited/Uneven Match
- Weak Match (match score = 0)

The methodology was specific in converting the “0” and “2” ratings among the four labels, but allowed for flexibility in converting the “1” rating. The methodology emphasized that professional judgment and in-depth discussion among the reviewers was required to determine whether a match score of “1” translates into a Good Match or a Limited/Uneven Match.

## ***Accessibility Review***

The evaluation methodology includes a review of the extent to which an assessment program's tests are fair to all students, including ELs and SWDs. Similar to evaluation of the other CCSSO criteria, this review also involves examining documentation (Generalizability criteria) specific to accessibility and exemplar items (Outcome criteria) that show how the program provides accessibility features and accommodations and/or item design that are fair while remaining valid assessments of the construct. Examining exemplar items provides evidence of what students actually experience while the documentation provides evidence of the program's rationales, research, design, development, and review processes.<sup>3</sup>

### ***Modifications from Prescribed Methodology***

When implementing the test content evaluation methodology in the current study, we were careful to adhere to the Center's guidance and specifications. However, as one of the first organizations to implement this innovative methodology, we encountered a number of situations that required us to modify the original methodology. We have already mentioned particular challenges with the Transparency review. This and other modifications we made from the prescribed methodology are described further in the main body of this report.

### ***Summary of Results***

The methodology specifies that, for each content area evaluated against the CCSSO criteria, an assessment program receive ratings for Content, Depth, and Accessibility. The Content rating provides evidence that the program assesses the content most needed for college and career readiness. The Depth rating provides evidence that the program assesses the depth that reflects the demands of college and career readiness. The Accessibility rating provides evidence that the program makes its assessments accessible to all students including ELs and SWDs.

Table ES2 presents a summary of the high school ratings for the four programs, followed by brief narrative descriptions of the Content, Depth, and Accessibility results. Additional information about these results can be found in Chapters 4 and 5 of this report.

---

<sup>3</sup> The Accessibility and Accommodations Manuals used to review the ACT Aspire, PARCC, and Smarter Balanced programs were dated 2015 and for the MCAS program the manual was dated 2014.




























Table ES2. Summary of Four Programs' High School ELA/Literacy and Mathematics Ratings

High School English Language Arts/Literacy				
Criteria	ACT Aspire	MCAS	PARCC	SBAC
<b>I. CONTENT: Assesses the <u>content</u> most needed for College and Career Readiness</b>				
<b>B.3 Reading:</b> Tests require students to read closely and use specific evidence from texts to obtain and defend correct responses. <sup>1</sup>				
<b>B.5 Writing:</b> Tasks require students to engage in close reading and analysis of texts. Across each grade band, tests include a balance of expository, persuasive/argument, and narrative writing.				
<b>B.6 Vocabulary and language skills:</b> Tests place sufficient emphasis on academic vocabulary and language conventions as used in real-world activities.				
<b>B.7 Research and inquiry:</b> Assessments require students to demonstrate the ability to find, process, synthesize and organize information from multiple sources.				
<b>B.8 Speaking and listening:</b> Over time, and as assessment advances allow, the assessments measure speaking and listening communication skills. <sup>2</sup>				
<b>II. DEPTH: Assesses the <u>depth</u> that reflects the demands of College and Career Readiness</b>				
<b>B.1 Text quality and types:</b> Tests include an aligned balance of high-quality literary and informational texts.				
<b>B.2 Complexity of texts:</b> Test passages are at appropriate levels of text complexity, increasing through the grades, and multiple forms of authentic, high-quality texts are used. <sup>3</sup>				
<b>B.4 Cognitive demand:</b> The distribution of cognitive demand for each grade level is sufficient to assess the depth and complexity of the standards.				
<b>B.9 High-quality items and variety of item types:</b> Items are of high technical and editorial quality and test forms include at least two items types, at least one that requires students to generate a response.				

Legend:

Excellent Match    
 Good Match    
 Limited/Uneven Match    
 Weak Match    
 Insufficient Evidence

**Table ES2. (Continued)**

<b>High School Mathematics</b>				
<b>Criteria</b>	<b>ACT Aspire</b>	<b>MCAS</b>	<b>PARCC</b>	<b>SBAC</b>
<b>I. CONTENT: Assesses the <u>content</u> most needed for College and Career Readiness</b>				
<b><u>C.1 Focus:</u></b> Tests focus strongly on the content most needed in each grade or course for success in later mathematics (i.e., Major Work).				
<b><u>C.2: Concepts, procedures, and applications:</u></b> Assessments place balanced emphasis on the measurement of conceptual understanding, fluency and procedural skill, and the application of mathematics.				
<b>II. DEPTH: Assesses the <u>depth</u> that reflects the demands of College and Career Readiness</b>				
<b><u>C.3 Connecting practice to content:</u></b> Test questions meaningfully connect mathematical practices and processes with mathematical content.		IE		
<b><u>C.4 Cognitive demand:</u></b> The distribution of cognitive demand for each grade level is sufficient to assess the depth and complexity of the standards.				
<b><u>C.5 High-quality items and variety of item types:</u></b> Items are of high technical and editorial quality and test forms include at least two item types, at least one that requires students to generate a response.				

**Legend:**

 <b>Excellent Match</b>	 <b>Good Match</b>	 <b>Limited/Uneven Match</b>	 <b>Weak Match</b>	 <b>Insufficient Evidence</b>
--	---	---	---	--

*Note.* MCAS = Massachusetts Comprehensive Assessment System; PARCC = Partnership for Assessment of Readiness for College and Careers; SBAC = Smarter Balanced Assessment Consortium.

<sup>1</sup> The criteria that are recommended to be more heavily emphasized have been underlined.

<sup>2</sup> Criterion B.8 is to be assessed over time and as advances allow; thus, the Criterion B.8 ratings were not considered when determining the composite Content rating (indicated by the gray shading).

<sup>3</sup> The Criterion B.2 rating is based solely on program documentation as reviewers were not able to rate the extent to which quantitative measures are used to place each text in a grade band. Thus, reviewers did not consider the Criterion B.2 rating as heavily when deciding the overall depth rating (indicated by the gray shading).

## **ELA/Literacy Content**

The composite ELA/literacy Content rating is based on five criteria: Close Reading (B.3), Writing (B.5), Vocabulary and Language Skills (B.6), Research and Inquiry (B.7), and Speaking and Listening (B.8). Across the four programs included in this study, ELA/literacy Content ratings ranged from Excellent to Weak. Assessments for all of the programs except ACT Aspire required students to read closely and use evidence from texts (Criterion B.3). The PARCC and Smarter Balanced assessments emphasized writing tasks that required students to engage in close reading and analysis of texts so that students can demonstrate college- and career-readiness abilities (Criterion B.5). The assessments for those same two programs also required students to demonstrate proficiency in the use of language, including vocabulary and conventions (Criterion B.6). All of the assessments except MCAS required students to demonstrate research and inquiry skills by finding, processing, synthesizing, organizing and using information from sources (Criterion B.7). Only the Smarter Balanced assessments currently assess listening skills; none of the programs assessed speaking skills at the time this study was implemented (Criterion B.8). The criteria acknowledge the need to assess speaking and listening skills, but they indicate this should be done over time and as assessment advances allow; the speaking and listening score does not contribute to the composite Content rating.

## **ELA/Literacy Depth**

The composite ELA/literacy Depth rating is based on four criteria: Text Quality and Types (B.1), Complexity of Texts (B.2), Cognitive Demand (B.4), and High Quality Items and a Variety of Item Types (B.9). The assessments for Smarter Balanced and ACT Aspire require students to demonstrate the range of thinking skills, including higher-order skills, while the MCAS and PARCC assessments require students to demonstrate less of a range. It should be noted that the PARCC assessments require *higher* cognitive demand than prescribed by the methodology while the MCAS assessment requires lower cognitive demand. For text quality and balance of types (Criterion B.1), the Smarter Balanced assessments received an Excellent Match rating while the ACT Aspire and MCAS received a Good Match rating, and the PARCC assessment received a Limited Match rating. All four programs' assessments required appropriate levels of text complexity and had multiple forms of authentic, previously published texts (Criterion B.2). The ACT Aspire and Smarter Balanced programs had assessments that required students to demonstrate a range of higher-order, analytical thinking skills in reading and writing based on the depth and complexity of college- and career-readiness standards, allowing robust information to be gathered for students with varied levels of achievement (Criterion B.4). All of the programs except ACT Aspire

had assessments comprising high-quality items as defined by the CCSSO criteria and that included a variety of item types strategically used to appropriately assess the standards (Criterion B.9).

### **Mathematics Content**

The composite mathematics Content rating is based on two criteria: Focus (C.1) and Concepts, Procedures, and Applications (C.2). All of the assessments except ACT Aspire focused strongly on the content most needed in high school for later success in mathematics (Criterion C.1). The PARCC and Smarter Balanced assessments measured conceptual understanding, fluency and procedural skill, and application of mathematics, as indicated in the college- and career-ready standards (Criterion C.2).

### **Mathematics Depth**

The composite mathematics Depth rating is based on three criteria: Connecting Practice to Content (C.3), Cognitive Demand (C.4), and High-Quality Items and a Variety of Item Types (C.5). Smarter Balanced, ACT Aspire, and PARCC fared well on Depth in mathematics while MCAS received a rating of Limited Match. All of the assessments except MCAS included brief questions and longer questions that connected the most important high school mathematical content to mathematical practices (Criterion C.3); insufficient information was provided for the MCAS program to determine the extent to which its assessment connected important content to mathematical practices. The PARCC and Smarter Balanced programs required students to demonstrate a range of higher-order analytical thinking skills and included questions, tasks, and prompts that measured basic and complex content intended by the college- and career-readiness standards (Criterion C.4). All of the programs except ACT Aspire had assessments comprising high-quality items as defined by the CCSSO criteria and that included a variety of item types strategically used to appropriately assess the standards (Criterion C.5).

### **Accessibility**

A narrative summary of the Accessibility results for the four programs is presented below. Given problems encountered in the review process, we do not provide Accessibility ratings. As previously noted, the Center’s test content evaluation methodology addresses Accessibility as a “light touch” review, with an emphasis on documentation and only a sample of exemplar items included in the evaluation. Ironically, even with such a “light touch” review concept, there was a voluminous amount of documentation provided to describe each program’s universal design and research underpinnings, assessment development processes, and accessibility and accommodation offerings. Reviewers were not always able to locate

sufficient information that was relevant to the sub-criterion to make fully informed match score ratings. This was particularly an issue because the accessibility rating criteria are very specific and stringent. The Center's forthcoming test characteristics methodology, that considers data from administered tests, will support a fuller examination of accessibility and presumably improve the usefulness of accessibility ratings.

A brief summary of accessibility reviewer comments on each program is provided next; each summary addresses program strengths followed by areas for improvement. Table 21 in the main report shows the accessibility features offered by each program.

### **ACT Aspire**

The ACT Aspire summative assessments are administered online or as a paper version, by each state's choice. The program provides a range of accessibility features and accommodations (e.g., eliminating irrelevant language demand, color contrast, limiting motor load, avoiding extraneous graphics), with similar accessibility features and accommodations offered for the paper-based and online assessments. Documentation includes a rationale for how each feature or accommodation supports valid score interpretations, when each may be used, and instructions for administration. ACT Aspire demonstrates strong adherence to universal design principles in its development of the assessed content areas. Information about the types of accommodations offered by ACT Aspire is available on their website as well as information about the type of student who might benefit from each based on best practices and research.

It was unclear how the program used information about the types of accommodations available and the type of student who might benefit from each when developing items and assembling forms. Also, the program's implementation of its universal design principles may not have been fully realized during item development and form assembly. For example, reviewers found documentation that indicated the program would provide multiple accommodations but within the documentation provided they were unable to find information about how the program would manage providing multiple accommodations for a single student.

### **MCAS**

The MCAS summative assessments are paper-based. The program offers standard accommodations that change the routine conditions under which a student takes the MCAS (e.g., frequent breaks, unlimited time, magnification, small group) and nonstandard accommodations (modifications) that change a portion of what the test is intended to measure (e.g., read aloud or scribe in ELA, calculator or non-calculator portions of mathematics). These accommodations are



provided to students with disabilities as determined by their Individualized Education Plan or 504 Plan and in accordance with the state's participation guidelines. In general, reviewers judged the accommodations and accessibility features offered by MCAS for its summative assessments to be reasonable. MCAS documentation reflected the program's efforts to consider universal design.

There were limited accommodations indicated specifically for ELs. (MCAS provided the *Requirements for the Participation of English Language Learners* after this study was completed that address, at least in part, deficiencies reviewers found.) Although reviewers judged the accommodations and accessibility features offered by MCAS to be reasonable, they also thought they were limited and did not maintain pace with the field. The program's use of universal design was perceived to be limited (based on the narrow populations considered and the limited feedback obtained during item development and bias reviews). The program offers a limited scope of accessibility features for some items and certain accommodations appear to introduce the opportunity for errors because student responses need to be transposed or items had to be skipped. Reviewers did not find a strong connection between research and the accommodations that MCAS made available in the provided documentation.

### **PARCC**

The PARCC summative assessments are administered online and offer paper-based assessments for students, as appropriate. The program incorporates accessibility features that are available to all students (e.g., color contrast, eliminate answer choices, highlight tool, pop-up glossary) and offers several test administration considerations for any student (e.g., small group testing, separate location, adaptive and specialized equipment or furniture), as determined by school-based teams. The program also offers a wide range of accommodations for SWDs (e.g., assistive technology, screen reader, Braille note-taker, word prediction external device, extended time) and ELs (e.g., word-to-word dictionary, speech-to-text for mathematics general directions provided in a student's native language, text-to-speech for the mathematics assessment in Spanish). PARCC was viewed favorably for its sensitivity to the design of item types that reflect individual needs of students with disabilities, and for its strong research base and inclusion of existing research on ELs. Reviewers found the accommodations offered by PARCC to be valid and appropriate based on current research.

Based on the information reviewed during the evaluation, reviewers were unable to locate information about the research needed to determine whether the accessibility features and accommodations that are offered by the program alter the constructs measured in its assessments. Specifically, reviewers noted that clearer documentation may be needed

regarding how PARCC administers multiple features simultaneously and the implications of how multiple accessibility features impact student performance. After the workshop, PARCC provided information about how they conduct trials and customer acceptance testing to ensure multiple features and embedded accommodations are properly working that addressed, at least in part, deficiencies that reviewers found.

### ***Smarter Balanced***

The Smarter Balanced summative assessments are administered online as adaptive tests while paper-based versions are offered as an accommodation. The program provides a range of accessibility resources: universal tools, designated supports, and accommodations. Depending on preference, students can select a number of universal tools that are embedded (e.g., digital notepad, highlighter, zoom, English glossary) or non-embedded (e.g., protractor, scratch paper, thesaurus, English glossary) within the assessment. The program also offers a number of designated supports to all students for whom the need has been indicated by an educator or team of educators. The designated supports can be embedded (e.g., color contrast, magnification, translations for the online version) or non-embedded (e.g., color contrast, separate setting, translations for the paper or online versions, translated glossary). For students with documented Individualized Education Plans or 504 Plans, several embedded accommodations are available, (i.e., American Sign Language, Braille, closed captioning, and text-to-speech) and several non-embedded accommodations (e.g., abacus, read aloud, scribe, speech-to-text) are offered. The program has specific guidelines for accessibility for ELs that highlight using clear and accessible language when developing items. Smarter Balanced's use of universal design and evidence-based design were described well. The program also appropriately suggests usability guidance to help educators support determinations of how different accommodations, designated supports and universal tools might interact.

The program's item development procedures incorporated accommodations and accessibility features from conception, which is consistent with the criteria. However, decision making rules were judged to be overly complicated and challenging for educators to apply. For SWDs, certain guidelines were judged to be overly prescriptive when there did not seem to be a reason for such strict guidance. After the workshop, Smarter Balanced highlighted the usability guidance that helps educators support determinations of appropriate accommodations, designated supports and/or universal tools and how they might interact in the *Individual Student Assessment Accessibility Profile* documentation. This information addressed, at least in part, deficiencies that reviewers noted.

### *Program Responses to Study*

We offered the programs included in this study the opportunity to comment about their participation, including commentary about the results relevant for their assessment and remarks about the test content evaluation methodology. We encouraged each program to include information in their response that might provide background and/or reasoning behind their test design and development as well as any other information that might help interpret this study's results regarding their assessments. The programs' responses are presented in Appendix K.

### *Conclusions*

Generally, implementing the methodology for the four programs went smoothly. However, there were a number of challenges that we and the reviewers experienced when implementing this evaluation methodology for the first time. This is not surprising, as it is not unusual to find that not every element works in practice as intended and that fine-tuning is needed. Moreover, any assessment review methodology needs to consider the desire of a comprehensive and in-depth review and the realistic constraints of time and other resources available to conduct the review. In general, the Center balanced efficiency and depth when developing its evaluation methodology.

Of note, the four programs that participated in the study made different choices about the design and specifications for their assessments (e.g., test design, coverage of the CCSS or other content standards). Some of these choices reflect operational considerations (e.g., administration time) that are not embedded in CCSSO's criteria, yet are reflected in the results obtained from the current study. Further, there are practical concerns such as testing time and cost that are not included in the criteria, but may be important assessment adoption considerations.

Detailed descriptions of the challenges we encountered when conducting this study, along with recommended revisions for future implementation of the test content evaluation methodology are described in Chapter 6 of this report.

# Evaluating the Content and Quality of Next Generation High School Assessments

## Chapter 1: Introduction

### *Background*

In the new era created by the federal *Every Student Succeeds Act* signed into law in December 2015, states have the responsibility to ensure their educational systems produce students who are prepared for the worlds of higher education and work. Student assessments are the primary mechanism for gauging the success of state educational systems. There are a variety of assessments that can be used, including those developed by individual states, consortia of states, and commercial vendors. Each of these assessments will have particular strengths and weaknesses that make them more or less suitable for specific applications. The question then becomes, how can policy-makers obtain clear, thorough, and unbiased characterizations of particular student assessments that reflect the complexities of next generation testing goals, strategies, and formats?

The National Center for the Improvement of Educational Assessment (NCIEA), hereafter referred to as the Center, developed an innovative evaluation methodology to address this complex question. The methodology goes well beyond traditional studies that examine the alignment between discrete test items and learning objectives. It takes as its guiding framework elements of the *Criteria for Procuring and Evaluating High Quality Assessments*, which was developed by the Council of Chief State School Officers (CCSSO) and released in 2014. CCSSO developed its criteria to be applicable to any assessment that was intended to measure college- and career-ready content standards in mathematics and English language arts (ELA)/literacy, especially the Common Core State Standards (CCSS).

### *Overview of Study*

In the present study, we used this new methodology to evaluate the extent to which the high school ELA/literacy and mathematics summative assessments for four programs match the CCSSO criteria relevant to test content. The following questions were addressed:

- Do the assessments place strong emphasis on the most important content of college and career readiness as called for by the CCSS and other college and career-readiness standards? (Content)

- Do the assessments require all students to demonstrate the range of thinking skills, including higher-order skills, called for by those standards? (Depth)
- Are the assessments accessible to all students, including ELs and SWDs? (Accessibility)

Included in this evaluation were summative (end-of-year) high school assessments developed by ACT Aspire, the state of Massachusetts, the Partnership for Assessment of Readiness for College and Careers (PARCC), and the Smarter Balanced Assessment Consortium (Smarter Balanced).

### ***Parallel Study***

A parallel study was conducted by the Thomas B. Fordham Institute (hereafter referred to as Fordham), which implemented the Center’s methodology for grades 5 and 8 summative mathematics and ELA/literacy assessments. Taken together, HumRRO and Fordham were first to implement the Center’s evaluation methodology. HumRRO and Fordham conducted their studies separately; however, the two organizations communicated often about the evaluation methodology and collaborated on the steps to implement it. To conserve time and resources, several activities were conducted jointly.

Fordham is publishing their findings for the grades 5 and 8 assessments in a separate report.<sup>4</sup> That report also offers contextual information that is useful for thinking about the results of these parallel studies while the present report focuses more on some of the methodological details.

### ***Participating Assessment Programs***

The four assessment programs included in the study—ACT Aspire, Massachusetts Comprehensive Assessment System (MCAS), PARCC, and Smarter Balanced—were intended to represent an array of options for states to consider for adoption. ACT Aspire is anchored by its capstone college readiness assessment, the ACT, a well-known college admissions test. MCAS is a highly regarded state developed and administered assessment. PARCC and Smarter Balanced are state membership-based consortia that were each federally funded to develop assessments based on the CCSS. Table 1 presents a summary of key characteristics of these four assessment programs, which vary considerably with regard to target content, test length, and format.

---

<sup>4</sup> Doorey, N., & Polikoff, M. (2016). *Evaluating the content and quality of next generation assessments*. Washington, DC: Thomas B. Fordham Institute.

**Table 1. Key Characteristics of the Four Assessments Included in the Evaluation Study**

Program	Subjects Reviewed	Mode	Grade Reviewed	Testing Time
ACT Aspire	Mathematics ELA/literacy - English - Reading - Writing	Online	Grade 10	3 hrs, 15 mins
2014 MCAS	Mathematics ELA/Literacy	Paper-Pencil	Grade 10	3 hrs, 30 mins
PARCC	Mathematics - Mathematics III - Algebra II ELA/Literacy	Online	Grade 11	7 hrs, 30 mins <sup>1</sup>
Smarter Balanced	Mathematics ELA/Literacy	Online, adaptive	Grade 11	5 hrs, 30 mins

<sup>1</sup> The 2015-2016 revisions will reduce this by an estimated one and one-half hours.

The **ACT Aspire** summative assessments are administered online (although a paper-based option is available) and were designed to measure the *ACT College and Career Readiness Standards*. The ACT Aspire summative assessments are administered to students in grades 9 and 10. The mathematics summative assessment is a single test while the ELA/literacy summative assessment comprises three separate tests—English, reading, and writing. The ACT Aspire assessments include multiple item types including selected-response items, constructed-response tasks, and technology-enhanced items and tasks. In 2015, the total testing time for the ACT Aspire grade 10 ELA/literacy and mathematics summative assessments was 3 hours and 15 minutes.

The **MCAS** summative assessments are administered via paper-pencil and are designed to measure the *Massachusetts Curriculum Framework* learning standards. Following adoption, Massachusetts modified the MCAS ELA/literacy and mathematics assessments to align with the CCSS. The MCAS grade 10 tests in ELA/literacy and mathematics were evaluated in this study. The MCAS assessments include multiple-choice items, short-answer questions, short-response items, open-response items, and writing prompts. The 2014 MCAS assessments were evaluated in the current study; the total testing time for the 2014 MCAS grade 10 ELA/literacy and mathematics summative assessments was 3 hours and 30 minutes.

The **PARCC** summative assessments were developed to measure the CCSS. These assessments are administered online and include two mandatory components. The grade 11 assessments were evaluated in this study. For mathematics, this translated to the Mathematics III and Algebra II assessments. The PARCC assessments have two components—a Performance-Based Assessment (PBA) intended to test students' ability to integrate and synthesize ideas from sources and to write (ELA/literacy) or complete multi-step, real world application problems (mathematics) and an End-of-Year Assessment (EOY) intended to test students' reading comprehension levels (ELA/literacy) or conceptual understanding (mathematics). This study included a review of both the PBA and the EOY. The PARCC assessments include multiple item types including constructed-response items, evidence-based selected response items (single correct response and multiple correct response), and technology-enhanced constructed response items. In 2015, the total testing time for the PARCC high school ELA/literacy and mathematics summative assessments was 7 hours to 7 hours and 30 minutes. PARCC intends to reduce the total time to 5–6 hours through revisions to its assessments.

The **Smarter Balanced** summative assessments are administered online and also consist of two parts—a computer adaptive test (CAT) and computer-based performance tasks. The Smarter Balanced summative assessments were developed to measure the CCSS. The grade 11 high school assessments were evaluated in this study. The Smarter Balanced assessments include multiple item types including multiple-choice (single correct response and multiple correct response); two-part multiple-choice, with evidence responses; matching tables; hot text; drag and drop; short text response; essay; hot spot; and short text and fill-in tables. In 2015, the total testing time for the Smarter Balanced high school ELA/literacy and mathematics summative assessments was 5 hours and 30 minutes.

### ***Organization of Report***

The following chapter (Chapter 2) provides an overview of the Center's evaluation methodology. Chapter 3 describes implementation details, including a description of how the experts who participated in the various workshops to review the assessment programs' documentation and test items were selected and trained. Chapter 4 presents the evaluation results pertaining to test content (what the methodology describes as Outcomes and Generalizability) and Chapter 5 presents the evaluation findings related to accessibility of the assessments to all prospective examinees. Chapter 6 provides commentary on some of the challenges in implementing this evaluation methodology as well as suggestions for improving both the efficiency and effectiveness of this very novel approach to test program evaluation.

## Chapter 2: Overview of Test Content Evaluation Methodology

The Center developed its new review methodology in an effort to determine the extent to which assessments that purport to measure college- and career-readiness standards are of high quality and sufficiently assess the more complex competencies needed for college and career readiness, including writing, research and inquiry, and higher-order, critical thinking skills. It is a tall order to conceive of a methodology that can do this in a way that accommodates different test designs, item types, and so forth, all while being mindful of resource requirements (e.g., development costs, testing time). The Center also sought to create a methodology that makes effective use of professional judgment and expresses evaluation findings in a way that is interpretable to all stakeholders and actionable by those responsible for the assessment program.

The Center was still developing its methodology at the time of our study, so the Center provided us with drafts of the scoring templates (ELA/literacy, mathematics, and accessibility), to guide our work. The Center recently published its methodology and it can be found at [http://www.nciea.org/publication\\_PDFs/Guide%20to%20Evaluating%20CCSSO%20Criteria%20Test%20Content%2001%2024%2016.pdf](http://www.nciea.org/publication_PDFs/Guide%20to%20Evaluating%20CCSSO%20Criteria%20Test%20Content%2001%2024%2016.pdf).

### *Study Criteria*

The CCSSO developed criteria for states and others to consider as they develop procurements and evaluate options for high quality summative assessments that purport to align to college and career readiness standards. Table 2 presents the subset of CCSSO criteria that were pertinent to the present study. The full set of CCSSO criteria is presented in Appendix A.

The Center's methodology translates the CCSSO criteria into specific rubrics and scoring procedures to facilitate both a credible and a practical evaluation of an assessment. To facilitate development of its methodology, the Center divided the CCSSO criteria into two parts: Test content and test characteristics. The test content evaluation procedures highlight the extent to which an assessment (a) aligns to content standards, (b) is accessible to all students, and (c) is transparent in its test design. The test characteristics evaluation procedures highlight the psychometric and statistical properties of an assessment, the quality of its administration, and how well the program reports and provides supplemental information to aid in interpreting and using test results to inform decisions. The focus of the present study is on the test content evaluation methodology and criteria, which is described further here.



**Table 2. Subset of CCSSO Criteria Implemented in Study of High School Assessments**

Criterion	Criterion Description
<b>Accessibility</b>	
A.5: Accessibility	Provide accessibility to all students, including ELs and SWDs
<b>Transparency</b>	
A.6: Transparency of test design and expectations	Assessment design documents and sample test questions are made publicly available
<b>ELA/Literacy</b>	
B.1: Text Quality and Types	Test forms include an aligned balance of high quality literary and informational texts
B.2: Complexity of Texts	Passages are at appropriate levels of text complexity, increasing thorough the grades, and multiple forms of authentic, high quality texts are used
B.3: Reading	Requires students to read closely and use specific evidence from texts to obtain and defend correct responses
B.4: Cognitive Demand	Distribution of cognitive demand for each grade level is sufficient to assess the depth and complexity of the standards
B.5: Writing	Requirements for students to engage in close reading and analysis of texts. Across grade band, tests include a balance of expository, persuasive/argumentative, and narrative writing
B.6: Vocabulary and Language Skills	Places sufficient emphasis on academic vocabulary and language conventions used in real-world activities
B.7: Research and Inquiry	Requires students to demonstrate the ability to find, process, synthesize, and organize information from multiple sources
B.8: Speaking and Listening	Over time and as advances allow, measures speaking and listening skills
B.9: High Quality Items and a Variety of Item Types	Items are of high technical and editorial quality and each test form includes at least two item types including at least one that requires students to generate a response.
<b>Mathematics</b>	
C.1: Focus	Test forms focus strongly on the content most needed in each grade or course for success in later mathematics (prerequisites for careers and a wide range of postsecondary studies)
C.2: Concepts, Procedures, and Applications	Places balanced emphasis on measurement of conceptual understanding, fluency and procedural skill, and application of mathematics
C.3: Connecting Practice to Content	Items meaningfully connect mathematical practices and processes with mathematical content
C.4: Cognitive Demand	Distribution of cognitive demand for each grade level is sufficient to assess the depth and complexity of the standards
C.5: High Quality Items and a Variety of Item Types	Items are of high technical and editorial quality and each test form includes at least two item types including at least one that requires students to generate a response

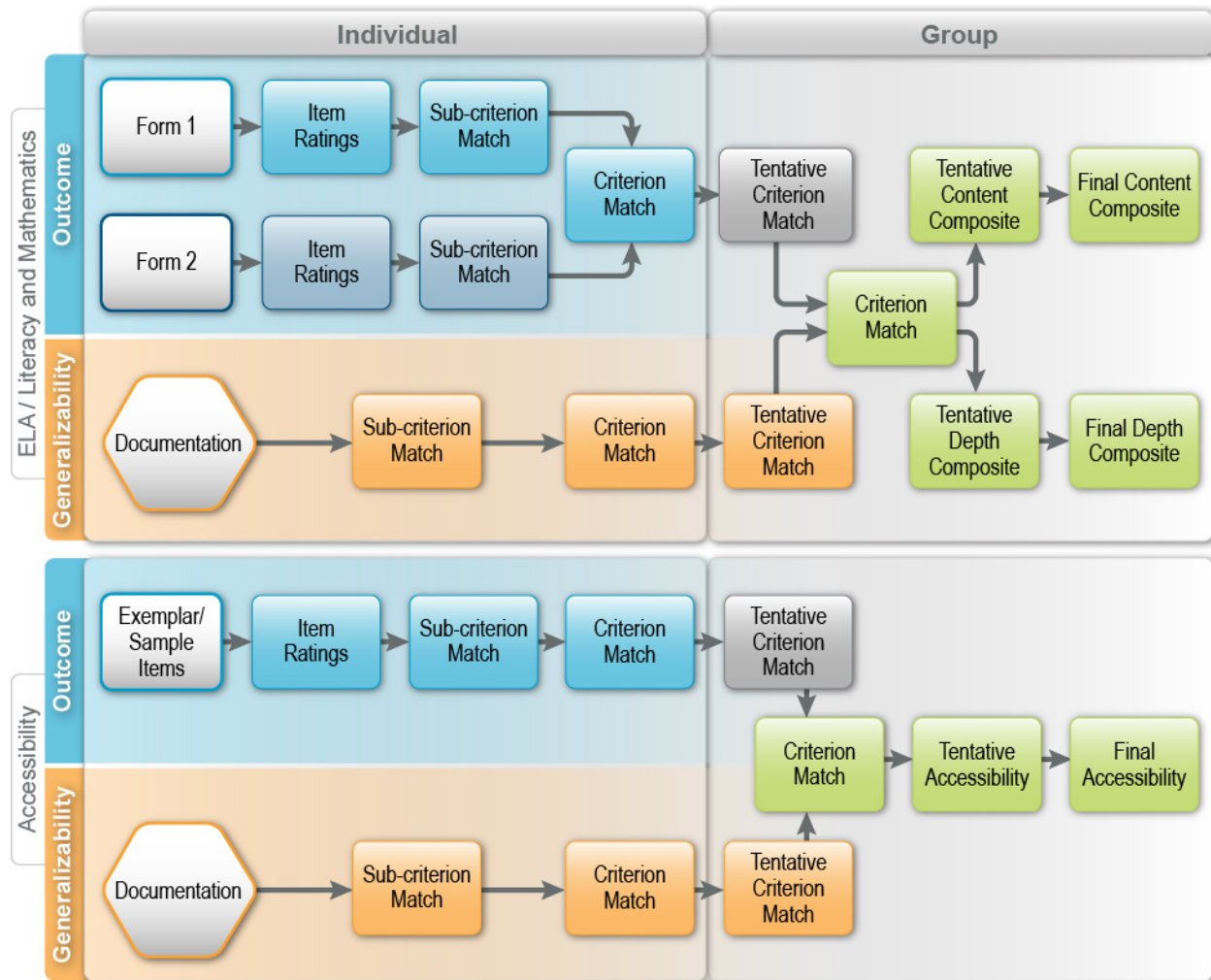
## *Review Process*

The methodology for evaluating assessments against the CCSSO test content criteria is a progressive process in which each step builds on the last step. To begin, individual subject matter experts review materials associated with the assessment program and judge the extent to which they match the applicable CCSSO sub-criteria. The ratings are successively aggregated culminating in two composites reflecting content and depth. In addition to ratings, reviewers at each stage of the process provide narrative comments that explain their ratings and highlight program strengths and weaknesses.

The methodology calls for the following four types of reviews:

- (1) Reviewing items and passages from two test forms to evaluate the extent to which operational assessment forms meet the scoring guidelines (Outcome review).
- (2) Reviewing test documentation and specifications to evaluate the extent to which results are generalizable across forms of the assessment (Generalizability review).
- (3) Reviewing the extent to which the assessment program provides sufficient information to the public regarding assessment design and expectations (Transparency review).
- (4) Reviewing the extent to which the assessment program's tests are fair to all students, including ELs and SWDs (Accessibility review).

The so-called Outcome and Generalizability reviews to evaluate test content and depth were conducted separately by HumRRO (for the high school assessments) and Fordham (for grades 5 and 8) and followed the Center's methodology quite closely. The Transparency review was conducted jointly by HumRRO and Fordham, and involved review of documentation that proved to be so voluminous that a suitable evaluation could not be conducted. The Accessibility review was conducted by HumRRO, with support from Fordham, and involved review of both Outcome (exemplar test items) and Generalizability (documentation). This review also proved particularly difficult to implement effectively. Figure ES1 provides an illustration of the methodology.



**Figure 1. Overview of rating process.**

A panel of experts individually reviews items from each of two forms of a test as the first step in the Outcome review. Those item ratings are summarized and used by individual reviewers to assign sub-criterion ratings (called match scores) for each form. Reviewers also record narrative comments that explain the basis for their individual match scores. The full group of Outcome reviewers then discusses their individual input and comes to consensus on (a) group match score ratings for each form, (b) group match score ratings for the program (i.e., across the two forms), and (c) narrative comments to provide further explanatory detail to each match score rating. Likewise, reviewers make independent match score ratings (along with an explanatory narrative) based on the test documentation (Generalizability) sub-criteria, then discuss as a group and derive a group sub-criterion rating and associated narrative comments. The last phase requires reviewers to discuss their group Outcome (item) and Generalizability (documentation) sub-criterion judgments and come to consensus on criterion-level ratings and narrative comments.

The final step involves aggregating the group criterion match scores to yield composite Content and Depth ratings for the program. The Outcome and Generalizability review panels could include the same, overlapping, or entirely different reviewers.

The Transparency review evaluates the extent to which various information about the tests are made publicly available. Other than using different criteria, the process is parallel to that described above for the Generalizability review. Since it involves review of the same program documentation, the Transparency review can be efficiently conducted by the same review panel that conducts the Generalizability review.

The Accessibility review involves review of exemplar items and documentation, so the process includes both Outcome and Generalizability reviews as described above. Because the materials reviewed are different from those reviewed by the aforementioned Outcome and Generalizability panelists, the Accessibility panelists form a distinct group. There also are some differences in the procedural details between reviewing the ELA/literacy and mathematics for content and quality, and the Accessibility reviews that will be described further in the next chapter.

### *Review Materials*

The test content evaluation methodology calls for reviewing the items (or performance tasks for more complex problem types) from two forms of the assessment. The forms should be operational and representative of the program's blueprints and other specifications and not be specially-created forms. Items on the forms are to be accompanied by metadata, such as scoring details and estimated cognitive complexity that will facilitate review by independent subject matter experts. This material is used for the Outcome review.

To gain an understanding of broader aspects of the assessment program, reviewers also evaluate documentation about the program, including information drawn from technical reports, specifications, websites, and other sources related to test design, development, administration, scoring, and maintenance practices. This documentation is used for the Generalizability and Transparency reviews.

### *Reviewers*

The methodology provides some flexibility in the make-up and number of individuals who comprise the review panels. Reviewers should be educators and/or experts in ELA/literacy, mathematics, large-scale assessment, and/or item writing. To the extent the reviewers possess the requisite expertise (e.g., knowledge of good item writing techniques, knowledge of accommodation and accessibility features) they can participate in one or multiple reviews. The

methodology permits a larger number of individuals to review operational items while fewer individuals review program documentation.

The methodology recommends that review panels comprise 5–8 members who represent a balance of knowledge and expertise for each content area and grade band. Grade span panels should reflect a range of characteristics, with at least three members possessing deep content expertise; it is expected that a person with content expertise might also have knowledge of the CCSS, assessment practices, and/or instruction at particular grade levels.

### **Rating System**

The Center’s methodology uses a 3-point rating system (0, 1, 2) to reflect the match between the CCSSO criteria and the material reviewed. These “match scores” are accompanied by narrative comments provided by the reviewers that further characterize and explain their judgments. Based on the methodology, determination of the group match scores and development of the narrative comments should include some individuals who participated in both the Outcome (item) and Generalizability (documentation) review.

Each group-level evaluation (sub-criterion, criterion, and composite) is converted into one of the following labels to characterize the match to the CCSSO criteria:

- Excellent Match (match score = 2)
- Good Match
- Limited/Uneven Match
- Weak Match (match score = 0)

Throughout the rating process, reviewers are encouraged to use their professional judgment. That said, the methodology is pretty straightforward about how to convert the “0” and “2” ratings among the four labels while allowing for more flexibility when converting the “1” rating. When converting this rating, the methodology emphasizes that an in-depth discussion among the reviewers is likely needed to determine whether a match score of “1” translates into a Good Match or a Limited/Uneven Match. While again recognizing the importance of professional judgment, the methodology provides specific guidance about weighting more heavily the criteria most important to college and career

#### **Evaluation Scores/Indicators**

- Match Score (0, 1, or 2)
- Narrative Comment
- Label (Excellent, Good, Limited/Uneven, Weak)

#### **Scored Elements**

- Sub-criteria
- Criteria
- Composites (Content, Depth)
- Accessibility

readiness. For example, the decision rules for determining the ELA/literacy composite Content rating specifies that criteria B.3 (close reading) and B.5 (writing) should be weighted more heavily than criteria B.6 (vocabulary and language skills), B.7 (research and inquiry), and B.8 (speaking and listening).

### ***Accessibility Review***

The evaluation methodology includes a review of the extent to which an assessment program's tests are fair to all students, including ELs and SWDs. This includes providing appropriate accommodations that will reduce construct-irrelevant variance while supporting valid interpretations of students' scores on the construct(s) being assessed. Programs might also provide features that are not formal accommodations that facilitate access to their assessments. Further, for the online assessments, the programs include universal design features that all students can access. Similar to evaluation of the other CCSSO criteria, this review also involves examining documentation (Generalizability criteria) specific to accessibility and exemplar items (Outcome criteria) that show how the program provides accessibility features and accommodations and/or item design that are fair while remaining valid assessments of the construct. Examining exemplar items provides evidence of what students actually experience while the documentation provides evidence of the program's rationales, research, design, development, and review processes.<sup>5</sup>

When developing the methodology, the Center acknowledged that an assessment program would likely have more accessibility features and/or accommodations than could practically be reviewed to evaluate the Accessibility criteria. Thus, the methodology associated with the Accessibility criteria calls for a "light touch" review, with each assessment program providing a limited sample of exemplar items with accompanying documentation. The Accessibility review provides important information about how the assessment program has considered making its assessment accessible to all students, especially ELs and SWDs. Because the Accessibility review does not examine all features and accommodations offered by an assessment program, the resulting information should be considered with the "light touch" intent in which it was conducted. The Center's test characteristics methodology, that considers data from administered tests, will support a fuller examination of accessibility.

### ***Modifications from Prescribed Methodology***

When implementing the test content evaluation methodology in the current study, we were careful to adhere to the Center's guidance and specifications. However, as one of the first

---

<sup>5</sup> The Accessibility and Accommodations Manuals used to review the ACT Aspire, PARCC, and Smarter Balanced programs were dated 2015 and for the MCAS program the manual was dated 2014.

organizations to implement this innovative methodology, we encountered a number of situations that required us to modify the original methodology. In some cases, the situation allowed us to implement the same or similar modification as did the Fordham team. These situations are described briefly below; additional details about the modifications that we implemented are provided in the Fordham report:

- **Challenges Associated with Particular Testing Programs:** The Center’s methodology was designed to be used to evaluate any assessment and, therefore, it did not perfectly fit each of the four program’s test design or specifications.
- **Cognitive Demand:** The assessments sometimes received low ratings because they required higher average cognitive demand rather than *matched* the demand of the standards.
- **Text Complexity Metadata:** To evaluate the criteria for text complexity, reviewers needed access to the metadata; however, we were unable to include text complexity data in our study because the programs often used different methods to evaluate text complexity, qualitative text complexity data varied across programs, and text complexity data were often too voluminous to display in the coding worksheet in any readable format.
- **Major Work of the Grade in Mathematics:** The relevant criterion uses the language of focusing *exclusively* on the major work of the grade, which would penalize items that mostly focus on major work yet include some non-major work content.
- **Weighing Criteria for Content and Depth Ratings:** The methodology recommends that certain criteria be emphasized more heavily when determining the composite (Content and Depth) ratings.

In contrast to Fordham, we did not encounter problems with the methodology related to item alignment and item quality (Criteria B.9 and C.5, High quality items and a variety of item types; refer to Table 1 for the full description of these criteria). Reviewers of the high school assessments believed that item-standard(s) alignment is an aspect of item quality and, therefore, we did not need to modify the methodology to evaluate these criteria. We also did not need to modify the methodology related to a balance among items that assess conceptual understanding, procedural skill and fluency, and application. Reviewers of the high school mathematics assessments were able to categorize items by their predominant focus (i.e., conceptual

understanding, procedural skill and fluency, and application) and determine if there was an adequate balance among the categories (Criterion C.2, Concepts, procedures, and applications).

As noted earlier, this study originally included an evaluation of test program transparency, or the extent to which programs provide sufficient information to the public regarding assessment design and expectations (Criterion A.6). Reviewers were challenged during the review of program documentation related to this criterion because of the vast amount of materials provided by the programs and the materials that are publicly available. Additionally, many testing programs continued to release additional information (such as sample items) after our review occurred, rendering this panel's findings somewhat outdated. Because of these challenges, we modified the methodology by dropping this criterion from our study.

We implemented two modifications to the methodology related to the Accessibility review. First, the Center's guidance recommended having panels review the program documentation and exemplar items separately by content area (ELA/literacy and mathematics). However, due to the timing of the in-person Accessibility review and the availability of qualified reviewers, we convened a single panel of experts who reviewed the documentation and exemplar items for both ELA/literacy and mathematics. Because both content areas were reviewed by the same reviewers, we ensured each panel included at least one reviewer with ELA/literacy content knowledge, at least one reviewer with mathematics content knowledge, at least one reviewer with expertise in accommodations for ELs and SWDs, and at least one reviewer with expertise in universal design. Second, given the "light touch" review of accommodations and accessibility features, we provide the evaluation results using summary statements for Accessibility but not ratings.



## Chapter 3: Implementation of the Review Process

This chapter describes the procedures used when implementing the Center’s test content methodology to evaluate the high school assessments of the four testing programs.

### *Evaluation Criteria*

The Center operationalized the CCSSO criteria in its methodology by outlining the evidence and scoring guidance that should be reviewed and considered when making decisions about the quality of assessments students will take. Using Criterion C.1 as an example, Table 3 illustrates how each criterion is divided into various sub-criteria. This table also presents the associated scoring guidance to evaluate each sub-criterion. The ELA/literacy Scoring Template used for this study is presented in Appendix B and lists the full set of ELA/literacy criteria and sub-criteria, evidence descriptors, location of evidence, scoring guidance, and tentative cut-offs. The Mathematics and Accessibility Scoring Templates followed the same pattern and are presented in Appendices C and D, respectively. The tentative scoring cut-offs included in Appendix D were developed by the Center specifically for this study.<sup>6</sup>

***Table 3. Illustration of Criterion C.1 with Sub-Criteria, Scoring Guidance, and Tentative Cut-Off***

Sub-Criterion	Scoring Guidance	Tentative Cut-Off
<b><i>C.1: Focusing Strongly on the Content Most Needed for Success in Later Mathematics</i></b>		
C.1.1: Most Important Content Assessed	<p>Calculate the percentage of score points that assess the most important content. Assign a score and provide notes under Comments (for each form).</p> <p>For High School:            2 – Meets: At least half of the score points in each course or grade align exclusively to prerequisites for careers and a wide range of postsecondary studies and all or nearly all domains within the widely applicable prerequisites are assessed.</p> <p>1 – Partially Meets: Nearly half of the score points in each course or grade align exclusively to prerequisites for careers and a wide range of postsecondary studies and the large majority of domains within the widely applicable prerequisites are assessed.</p> <p>0 – Does Not Meet: Less than half of the score points in each course or grade align exclusively to prerequisites for careers and a wide range of postsecondary studies and/or less than the large majority of domains within the widely applicable prerequisites are assessed.</p>	<p>For High School:            2 – Meets: 50-100% of the score points align exclusively to the widely applicable prerequisites and/or at least 90% of the domains within the widely applicable prerequisites are assessed.</p> <p>1 – Partially Meets: 40-50% of the score points align exclusively to the widely applicable prerequisites and at least 75% of the domains are assessed.</p> <p>0 – Does Not Meet: 0-39% of the score points aligns to the major work and/or less than 75% of the domains are assessed.</p>

<sup>6</sup> The final Accessibility Scoring Template published by the Center does not include tentative cut-offs.

**Table 3. (Continued)**

Sub-Criterion	Scoring Guidance	Tentative Cut-Off
<b>C.1: Focusing Strongly on the Content Most Needed for Success in Later Mathematics</b>		
C.1.2: Assessment Design Reflect Important Content	<p>Rate the extent to which the percentage of score points that assess the most important content is indicated in the specifications. Assign a score and provide notes under Comments:</p> <p>For High School:</p> <p>2 – Meets: The test blueprints or other documents indicate that at least half of score points in each course or grade align exclusively to prerequisites for careers and a wide range of postsecondary studies and all or nearly all domains within the widely applicable prerequisites are assessed.</p> <p>1 – Partially Meets: The test blueprints or other documents indicate that nearly half of score points in each course or grade align exclusively to prerequisites for careers and a wide range of postsecondary studies and the large majority of domains within the widely applicable prerequisites are assessed.</p> <p>0 – Does Not Meet: The test blueprints or other documents indicate that less than half of score points in each course or grade align exclusively to prerequisites for careers and a wide range of postsecondary studies and/or less than the large majority of the domains within the widely applicable prerequisites are assessed.</p>	<p>For High School:</p> <p>2 – Meets: 50-100% of the score points align exclusively to the Major Work and/or less than 75% of the domains within the widely applicable prerequisites are assessed.</p> <p>1 – Partially Meets: 40-50% of the score points align exclusively to the Major Work and at least 75% of the domains are assessed.</p> <p>0 – Does Not Meet: 0-39% of the score points aligns to the Major Work and/or less than 75% of the domains are assessed.</p>

### **Review Materials**

To gather the evidence needed to implement the Center’s evaluation methodology, a variety of materials were needed, including test forms, metadata for items and passages, exemplars of accommodations and access features, and program documentation. Programs provided the requisite materials to ensure the most appropriate and current materials were included in the review. Each program identified a liaison with whom we worked to collect and organize the materials.

### **Test Forms**

The goal was to review two high school test forms each in ELA/literacy and mathematics for all four assessment programs. Reviewers evaluated two assessment forms for each assessment program with the exception of MCAS, which had only one assessment for each content area. With the exception of Smarter Balanced, programs were free to submit any two

operationally administered forms/events. Fixed forms were evaluated for ACT Aspire, MCAS, and PARCC.

Smarter Balanced assessments are computer adaptive, which means that the test adapts to the student's ability level such that subsequent items are selected based on how the student responds to a given item or cluster of items. Because the test adapts to the student's ability and is not a fixed form, Smarter Balanced refers to these as test events. Smarter Balanced summative assessments include an adaptive component and a performance task (PT) component. A set of PTs are assigned to a school, and within the set, a single PT is randomly assigned to an individual student. Generally, the CAT portions of the test adapt at the item level. However, items associated with reading and listening passages in the ELA/literacy tests are administered as a unit after the passage is chosen. A passage is selected based on the degree to which the set of items associated with the passage matches a student's ability. In addition, the selection of items and passages are constrained based on content requirements as described in the blueprint. This approach uses fewer items and allows scores to be produced with smaller margins of error for students who perform at the ends of the performance spectrum. It also provides the opportunity for greater variation in content across the test events. For its CAT, Smarter Balanced drew one test event from the events that were/could be administered to students at the 40<sup>th</sup> percentile of student achievement and drew the other test event from the events that were/could be administered to students at the 60<sup>th</sup> percentile of student achievement. Both test events were generated using the operational item selection algorithm. To evaluate the degree to which other forms would vary, the study also included review of the results of a simulation study of 1,000 forms per grade and content area as recommended by the Center's test content evaluation methodology.

### ***Metadata for Items and Passages***

The evaluation methodology requires the evaluation of specific metadata related to the items (ELA/literacy and mathematics) and passages (ELA/literacy). The programs provided all the requested metadata (listed in Appendix E); however, this was complicated by the fact that not all metadata were routinely captured by each program. To the extent possible, we pre-populated these metadata into customized electronic (Excel) rating forms so they would be readily accessible to reviewers. Due to security concerns raised by the programs, the keyed correct answer was not recorded in the Excel rating workbooks but rather reviewers were informed of the correct answer upon their request.

## ***Program Documentation***

For many programs, the documentation related to their assessments was voluminous. For the reviewers to conduct their review of the program documentation in an efficient manner, it needed to be organized and specific information related to each criterion highlighted. Specifically, the information was organized into tables unique to each program with columns that identified the information by criterion and/or sub-criterion, document name, location of relevant information within the document, and helpful notes for finding or interpreting the information. Each of the four programs prepared and/or reviewed the organized and highlighted information relevant to each criterion to ensure it was the most useful and current information available for reviewers to use when conducting their evaluation.

## ***Exemplars of Accommodations/Access Features and Fairness***

As part of the Accessibility review, programs provided exemplar items for both ELA/literacy and mathematics that incorporate some of their accommodations and access features so that reviewers might gain a better understanding of the program's handling of them. Documentation of a program's accommodations/access features and fairness were prepared and reviewed similarly to the documentation regarding other program features as discussed above.

Programs selected sets of exemplars that showed how their accommodations/access features and/or item design is fair for test-takers and support valid score interpretations. Each program selected at least one set of exemplar items for ELA/literacy and one set for mathematics. Each set was to consist of at least 10 but no more than 25 exemplar items; the exemplar items were to be accompanied by annotated descriptions of what the accommodation/access feature was and other helpful information (e.g., instructions for use). Programs submitted at least five exemplar items that provide accommodations or accessibility features for high incidence disabilities. A high incidence disability is one that is more common (e.g., speech and language impairment, learning disability, emotional disturbance) and occurs in about 1 in 10 school-age children.<sup>7</sup> Programs were permitted to provide at least one exemplar item for each usage that was essential for a particular disability. If the accessibility feature or accommodation was available only through a technology platform, the program was to provide instructions for how to locate and use the accessibility feature or accommodation in the same way a student would experience it while allowing the reviewer to explore it for the purposes of evaluation.

---

<sup>7</sup> <http://www.enotes.com/research-starters/low-incidence-high-incidence-disabilities>

## *Reviewer Selection*

The Center’s test content evaluation methodology relies heavily on expert judgment; therefore, we paid careful attention to ensuring that highly qualified, yet unbiased individuals served as reviewers in the study. First, we gathered reviewer recommendations from the four assessment programs that participated in the study, and from national assessment and content experts. Highly qualified individuals included those with content knowledge in ELA/literacy or mathematics, experienced classroom teachers, individuals with expertise in large-scale assessment, and experts in accommodating SWDs and ELs. Individuals who had served in an advisory role to an assessment program or served on the CCSS writing teams were considered eligible. Individuals who were or had been employed by the assessment program were not considered eligible to participate in the study.

Each potential reviewer completed a short application form detailing his or her relevant experience and potential conflicts of interest. After reviewing their qualifications, we selected 20 individuals to serve as Outcome reviewers (10 ELA/literacy reviewers and 10 mathematics reviewers). We also selected four individuals to serve as Generalizability reviewers; all of the individuals we selected as Generalizability reviewers also served as Outcome reviewers. The Generalizability review was conducted jointly with Fordham, so they also selected four individuals for this review. Finally, we selected a different set of nine individuals to serve as Accessibility reviewers.<sup>8</sup> There was no opposition by any of the programs regarding the selected individuals and reviewer participation was confirmed. Brief descriptions that highlight the background and expertise of the individuals selected to participate in these reviews are presented in Appendix F.

## *Review Activities*

### *Outcome Review*

The Outcome review occurred as an in-person, 4-day workshop that involved evaluating a variety of characteristics associated with the operational items and passages administered by the four programs. Separate workshops were conducted for ELA/literacy and mathematics. Ten reviewers participated in each content area, for a total of 20 reviewers who provided ratings on the Outcome criteria. Across the nine ELA/literacy Outcome criteria, reviewers made 20 sub-criteria ratings and across the five mathematics Outcome criteria, reviewers made six sub-criteria ratings. Generally, each group consisted of five reviewers who reviewed two forms of the

---

<sup>8</sup> A tenth reviewer was selected for this review but she had a last minute conflict and was not able to participate.

assessments for two programs. Because evaluation of the ELA/literacy assessment involves more sub-criteria than mathematics, the ELA/literacy groups were split into three groups during the workshop to complete the review of all forms across the four programs.

Reviewers first made independent ratings for the various item attributes associated with each sub-criterion and recorded narrative comments to support each rating. As reviewers entered their independent ratings into the electronic rating forms,<sup>9</sup> each reviewer's data were summarized into item summary data (e.g., percent of items aligned to standards, percent of items at depth of knowledge [DOK] level 1). The rating spreadsheet then auto-calculated tentative sub-criterion scores (i.e., match scores with values 0, 1, or 2) that reviewers then individually determined whether to keep or to change based on their professional judgment and comments they had made when reviewing the items.

The Outcome review workshop began with several hours of training that included a thorough review of the relevant criteria, guidance for interpreting and applying the criteria, instructions for navigating the online assessments, and details for recording their ratings electronically. The training was conducted separately for the two content areas (ELA/literacy and mathematics). Following the formal training, the 10 reviewers for each content area were grouped into two panels of five reviewers each. One HumRRO staff member served as facilitator of each group. For calibration purposes, reviewers in the separate panels together discussed the first several items and decided on the criteria ratings as a group rather than individually. This process continued until facilitators and reviewers were confident the criteria were applied consistently and accurately. At that time, reviewers evaluated the remaining items independently.

The facilitator of each panel monitored the progress of the reviewers as they completed their independent ratings and provided clarification and/or additional training, as needed. As the reviewers discussed ratings during each stage, the facilitator encouraged the panel to focus on key aspects of the method's scoring guidance so that the final rating would not be confounded by other considerations.

A final activity completed by the reviewers was to prepare narrative comments for each criterion. To prepare these comments, the facilitator encouraged the reviewers to refer to the individual and group comments recorded for each criterion as well as the evaluation criteria. The facilitator provided guidance as needed to ensure the comments included information

---

<sup>9</sup> The rating forms were Excel workbooks that auto-calculated each reviewer's individual item-summary ratings and allowed reviewers to enter narrative comments.

recommended by the methodology as well as important factors the reviewers considered when establishing the final criterion ratings.

### ***Generalizability Review***

The Generalizability review involved an evaluation of each program's documentation that provided information about their summative assessments; separate reviews were conducted for ELA/literacy and mathematics.<sup>10</sup> The methodology requires that the programs highlight which documents address each sub-criterion. In an attempt to reduce the burden for the programs, we created a table that listed all relevant Generalizability criteria, organizing the documents according to whether they provided key information relevant to each criterion. For each relevant criterion, we listed the title of each program document as well as the exact location of the key information within the document (chapter, section, page) and any notes that might be helpful for the reviewer when accessing or evaluating the information. The MCAS, PARCC, and Smarter Balanced programs reviewed and approved the table of information we prepared; ACT Aspire created its own information table.

The reviews were conducted remotely, and all reviewers participated in a 2-hour training session prior to beginning the review. During training, reviewers were presented the Generalizability criteria, and guidelines for interpreting and applying them during their review of the programs' documentation. As the criteria are different for each content area and reviewers participated based on their content expertise, separate training sessions were conducted for reviewers of the ELA/literacy and mathematics assessments. Following the training, reviewers were provided the materials needed to complete their review, including the final table of documents, along with electronic copies of the documents or links for accessing them. Reviewers also were given step-by-step instructions for evaluating the program information against the various criteria and an electronic form to record their individual ratings. A total of 20 ratings were needed across the nine ELA/literacy criteria and a total of seven ratings were needed across the five mathematics criteria.

Completing the individual ratings was a self-paced activity; however, reviewers needed to complete them within 8 days following training. Upon receipt, we compiled the individual reviewer ratings and aggregated them according to the methodology to achieve a tentative group match score for each criterion. We presented the tentative criterion group match scores during a Web conferencing session where the reviewers were encouraged to discuss the tentative scores and come to consensus on the final rating for each criterion. Separate Web

---

<sup>10</sup> The Generalizability review was conducted jointly with the Thomas B. Fordham Institute, with Fordham taking the lead role.

conferencing sessions were held for reviewers to decide the final ELA/literacy and mathematics match scores. After the final group match scores were determined, reviewers worked together to generate narrative comments that explained and provided context for each rating.

### ***Accessibility Review***

We conducted the Accessibility review to evaluate the extent to which each program's assessment is accessible to all students, including ELs and SWDs.<sup>11</sup> As noted earlier, the Accessibility review included a review of exemplar or sample items (Outcome evidence) and program documentation related to universal design, accessibility features, and accommodations (Generalizability evidence).

All Accessibility reviewers attended a 2-hour Web conferencing orientation session, followed by a 2-day in-person workshop to review and evaluate the programs' documentation and exemplar/sample items. This review was conducted across content areas (ELA/literacy and mathematics) and grades (grades 5, 8, and high school).<sup>12</sup> A total of nine individuals participated in the Accessibility review, all who had expertise in universal design, accessibility features, and accommodations for ELs and SWDs, and/or content knowledge in ELA/literacy or mathematics and across the grade span.

The in-person Accessibility workshop began with a thorough review of the relevant criteria, guidance for interpreting and applying the criteria, instructions for navigating the online assessments, and details for electronically recording ratings. This training was operationalized by having the reviewers rate the MCAS program together as a group, which also served the important purpose of calibrating the reviewers on the methodology and scoring guidance. Following review of the MCAS documentation and exemplar/sample items as a large group, reviewers were assigned to one of three groups (with three reviewers per group) to evaluate one of the remaining assessments. Each group included one HumRRO staff who served as facilitator.

The Accessibility reviewers followed a process similar to that described above to review the program documentation and exemplar/sample items. Reviewers individually evaluated the program's documentation against the relevant criteria and determined individual (Generalizability) criterion match scores. For each Generalizability sub-criterion, the reviewers evaluated the documentation separately for ELs and SWDs and determined how well the

---

<sup>11</sup> The Accessibility review was conducted jointly with the Thomas B. Fordham Institute, with HumRRO taking the lead role.

<sup>12</sup> ACT and PARCC provided exemplars at grades 5, 8, and high school. Smarter Balanced provided samples for grades 4, 8, and high school. MCAS provided exemplars for their grade 10 test.



information met each. They also individually evaluated the program's exemplar/sample items against the relevant criteria and determined individual (Outcome) criterion match scores. Reviewers evaluated exemplars for construct integrity and ease of use for ELs and SWDs, as appropriate. In addition, they determined the extent to which the exemplars provided evidence that the program had implemented what they described in their documentation. Reviewers' ratings were aggregated to form tentative group criterion match scores; reviewers discussed each tentative match score until they reached consensus on a final group criterion match score. Following the determination of a rating (individual or group) reviewers recorded narrative comments to provide context and explanatory support.

Across the Accessibility criteria, reviewers made 36 Generalizability ratings (one set of 18 for ELs and another set of 18 for SWDs) and two Outcome ratings (one for ELs and one for SWDs).

### ***Access to Exemplar/Sample Items***

Access to the exemplar/sample items provided by the testing programs varied significantly, due in part to the fact that the online testing platforms used by ACT Aspire, PARCC, and Smarter Balanced each limited the extent to which reviewers could view and interact with each item and in every form in which it might be available (e.g., in a program with many accommodations an item might be available in multiple languages and in multiple formats such as braille, large print, text to speech, with glossary). ACT Aspire provided access to its items with extensive metadata that outlined how the item should be treated across all of the possible accommodation and accessibility features. For example, they provided information on all possible presentation, interaction/navigation, and response processes for each exemplar item they provided. However, the exemplar items could only be viewed as a default user with default accessibility and accommodation features. Because MCAS is a paper-pencil only assessment, this program provided artifacts (rather than items) and included an American Sign Language (ASL) CD, text-to-speech DVD, mathematics manipulatives (block and ruler), high school Spanish test version, and a large print test version. The exemplar items provided by PARCC allowed the reviewers to view sets of items under certain accommodated conditions (e.g., screen-reader, text-to-speech), but reviewers could not view the items with their full range of accommodations. Smarter Balanced provided reviewers access to their online sample items which reviewers viewed under various accommodation and accessibility conditions. However, these sample items were available to anyone and were not operational items administered to students, as prescribed by the methodology.

## Scoring Procedures

### **Individual Item, Sub-criterion, and Criterion Match Score Ratings**

Once reviewers completed their independent sub-criterion ratings, tentative individual criterion match score ratings were auto-calculated. Individual panelists reviewed their tentative individual criterion match score ratings, along with the narrative comments they had recorded for each sub-criterion and, using their professional judgment, determined a final individual criterion match score rating.<sup>13</sup> The reviewers' final individual criterion match score ratings were aggregated to form tentative group criterion ratings.

### **Group Criterion Match Score and Composite Content and Depth Ratings**

The next step in the methodology required reviewers to discuss the tentative group criterion match score ratings and decide on final group criterion match score ratings. To establish these final ratings, reviewers considered the Outcome tentative group criterion match score rating along with the Generalizability tentative group criterion match score rating. When determining the final group criterion match score ratings, the reviewers discussed the tentative group match score rating, referring back to their individual narrative comments and using their professional judgment. It is important to note that the methodology emphasizes that for the ELA/literacy and mathematics ratings the Outcome rating be considered more heavily than the Generalizability rating when establishing the final Content and Depth individual and group criterion match score ratings. However, for Accessibility, the methodology calls for the Generalizability rating to be considered more heavily than the Outcome rating due to the fact that a limited number of exemplars are reviewed.

As noted earlier, the methodology uses a 3-point rating system (0, 1, 2) to reflect the match between the CCSSO criteria and the material reviewed, which is converted into labels of Excellent Match, Good Match, Limited Match, and Weak Match. Ratings of Excellent (2) and Weak (0) are relatively straightforward and likely require minimal discussion among reviewers, although they are encouraged to use their professional judgment to determine the final rating. Determining a match score rating of "1" (Good Match or Limited Match) requires professional judgment as well as in-depth discussion among the reviewers to determine whether that rating translates into a Good Match or a Limited/Uneven Match. The methodology does not require total agreement among the reviewers, but the final criterion rating should be the result of the reviewers

---

<sup>13</sup> The scoring templates provided tentative cut-offs for reviewers to interpret the more general language of the criterion; however, reviewers could and did use their professional judgment as they interpreted and applied the criteria.

coming to consensus on the most appropriate rating. Once the reviewers determined a final rating for each criterion, they recorded narrative comments to support and explain their rating, noting any minority views of the reviewers.

Once final group criterion ratings were determined, tentative composite Content and Depth ratings were generated. Using the same process as for the final group criterion match score ratings, the reviewers discussed separately the tentative criterion match score ratings that comprised the Content composite and the tentative criterion match score ratings that comprised the Depth composite. Referring to the group final criterion match score ratings and narrative comments for those criteria, the reviewers again used their professional judgment to determine final composite Content and final composite Depth ratings. Finally, the reviewers recorded narrative comments to support and explain each rating.

### ***Development of Summary Statements***

The final step in the evaluation methodology involved developing summary statements that provide context that help to interpret and support the final ratings. These statements were developed separately by program for each criterion, each composite, and overall. Following the in-person review workshop, project staff developed the summary statements using the final criterion and final composite ratings and the reviewers' narrative comments. The reviewers' narrative comments formed the bases of these statements; however, we added information from the criterion definition and/or scoring guidance so that individuals not familiar with the criteria would be better able to interpret the ratings. We also revised the narrative comments to ensure correct grammar and complete sentences. There was no attempt to make the contents of the comments parallel across the four assessment programs (ELA/literacy, mathematics, and accessibility).

## Chapter 4: Test Content and Depth Results

The methodology specifies that, for each content area evaluated against the CCSSO criteria, an assessment program receive ratings for Content, Depth, and Accessibility.<sup>14</sup> The Content rating provides evidence that the program assesses the content most needed for college and career readiness. The Depth rating provides evidence that the program assesses the depth that reflects the demands of college and career readiness. The Accessibility information provides evidence that the program makes its assessments accessible to all students, including ELs and SWDs; the Accessibility results for the four programs are presented in Chapter 5.

Presented below are Content and Depth results when implementing the test content methodology to evaluate the high school ELA/literacy and mathematics summative assessments for the four programs. Results are presented by content area, first for ELA/literacy followed by mathematics. Within each content area, results are presented overall by Content and Depth, followed by a brief summary of how the programs performed on the various criteria related to Content and Depth. Complete ratings and summary statements for ACT Aspire, MCAS, PARCC, and Smarter Balanced are presented in Appendices G – J, respectively.

### ELA/Literacy

#### Content

The composite ELA/literacy Content rating is based on five criteria: Close Reading (B.3), Writing (B.5), Vocabulary and Language Skills (B.6), Research and Inquiry (B.7), and Speaking and Listening (B.8). As can be seen in Table 4, PARCC and Smarter Balanced received the highest ratings (Excellent Match) and were judged to place a strong emphasis on the most important content for college and career readiness, while MCAS (Limited Match) and ACT Aspire (Weak Match) were judged to provide less of an emphasis on the most important content. An overall summary of how the programs performed on the Content criteria is provided below.

**Table 4. Composite ELA/Literacy Content Ratings**

ACT Aspire	2014 MCAS	PARCC	Smarter Balanced
			

<sup>14</sup> Recall this study originally included an evaluation of test program transparency, or the extent to which programs provide sufficient information to the public regarding assessment design and expectations (CCSSO Criterion A.6). However, due to several challenges associated with this review, we dropped this criterion from our study. These challenges are discussed in Chapter 6 of this report.

### Close Reading (Criterion B.3)

Criterion B.3 examines the extent to which the assessment requires students to read closely and use evidence from texts to obtain and defend responses. The assessment must include the following to fully meet this criterion:

- Nearly all reading items require close reading and analysis of text rather than skimming, recall, or simple recognition of paraphrased text.
- Nearly all reading items focus on central ideas and important particulars.
- Nearly all items are aligned to the specifics of the standards.
- More than half of reading score points are based on items that require direct use of textual evidence.

Table 5 presents how the four programs fared in meeting the requirements for this criterion.

**Table 5. Rating for Close Reading (Criterion B.3)**

ACT Aspire	2014 MCAS	PARCC	Smarter Balanced
W	G	E	E

Of the four programs, PARCC and Smarter Balanced assessments excelled on this criterion. Items on the assessments for these two programs focused on central ideas, aligned well to key aspects of the standards, and required students to read the text closely to find meaning and provide a response. Additionally, more than half the score points for these two programs were based on items that required the direct use of textual evidence. The MCAS program performed well on this criterion; however, these assessments fell short in requiring students to cite textual evidence in their response. The ACT Aspire assessments did not meet many of the requirements of this criterion.

### Writing (Criterion B.5)

Criterion B.5 examines the extent to which the assessment requires students to write narrative, expository, and persuasive/argumentative essays (across each grade band, if not in each grade) in which they use evidence from sources to support their claims. The assessment must include the following to fully meet this criterion:

- All three writing types are appropriately equally represented across all forms in the grade band (K–5; 6–12), allowing blended types (i.e., writing types that blend two or more of narrative, expository, and persuasive/argumentative) to contribute to the distribution.

- All writing prompts require writing to sources (meaning they are text-based).

Table 6 presents how the four programs fared in meeting the requirements for this criterion.

**Table 6. Rating for Writing (Criterion B.5)**

ACT Aspire	2014 MCAS	PARCC	Smarter Balanced
W	W	E	E

The PARCC and Smarter Balanced programs also excelled on this criterion. The writing prompts for these programs' high school assessments required students to write to textual sources and they included narrative, expository, and persuasive/argumentative writing types and/or a blended combination of two writing types. The ACT Aspire and MCAS assessments only included one writing type and the majority of items that assessed writing standards was multiple choice and did not require students to actually generate a written response. The MCAS assessment also included only one writing prompt and the prompt required students to write about a previously read passage, but it did not require the response to cite direct textual evidence.

### ***Vocabulary and Language Skills (Criterion B.6)***

Criterion B.6 examines the extent to which the assessment requires students to demonstrate proficiency in the use of language, including academic vocabulary and language conventions, through tasks that mirror real-world activities. The assessment must include the following to fully meet this criterion:

- The large majority of vocabulary items (i.e., three-quarters or more) focus on Tier 2 words and require the use of context, and more than half assess words important to central ideas. According to the standards, Tier 2 words are “general academic” words that are far more likely to appear in written text than in speech.
- A large majority (i.e., three-quarters or more) of the items in the language skills component and/or scored with a writing rubric (i.e., points in writing tasks that are allocated toward a language sub-score), mirror real-world activities, focus on common errors, and emphasize the conventions most important for readiness.
- Vocabulary is reported as a sub-score or at least 13% of score points are devoted to assessing vocabulary/language.
- Language is reported as a sub-score or at least 13% of score points are devoted to assessing language skills (language skills items plus score points).

Table 7 presents how the four programs fared in meeting the requirements for this criterion.

**Table 7. Rating for Vocabulary and Language Skills (Criterion B.6)**

ACT Aspire	2014 MCAS	PARCC	Smarter Balanced
L	L	E	E

The PARCC and Smarter Balanced programs both performed very well on this criterion. The large majority of vocabulary items on both of these high school assessments focused on Tier 2 words and required students to use context to determine meaning. Additionally, the large majority of items that measured language skills emphasized the conventions most important for readiness and mirrored real world skills and tasks. These assessments also reported vocabulary and language skills as sub-scores or devoted at least 13% of score points to assessing these skills. The ACT Aspire assessment used very few Tier 2 words to assess vocabulary and, although the majority of items required students to use context to determine meaning, most did not assess words important to central ideas. Additionally, ACT Aspire reported language skills as a sub-score, but this assessment did not report vocabulary as a sub-score nor did it devote sufficient score points to assessing vocabulary. The large majority of items on the MCAS high school assessment that assessed vocabulary used Tier 2 words; however, not all items required students to reference the text for context or meaning. Additionally, less than half of the items that assessed language skills mirrored real world activities. Finally, although language skills were reported as a sub-score, vocabulary was not, nor did the assessment devote sufficient score points to assessing vocabulary.

**Research and Inquiry (Criterion B.7)**

Criterion B.7 examines the extent to which the assessment requires students to demonstrate research skills, including the ability to analyze, synthesize, organize, and use information from sources. The assessment must include the following to fully meet this criterion:

- Three-quarters or more of the research items on each test form require analysis, synthesis, and/or organization of information.

Table 8 presents how the four programs fared in meeting the requirements for this criterion.

**Table 8. Rating for Research and Inquiry (Criterion B.7)**

ACT Aspire	2014 MCAS	PARCC	Smarter Balanced
G	W	E	E

The ACT Aspire, PARCC, and Smarter Balanced programs do a good job of including research items on their assessments that require students to analyze, synthesize, and/or organize information. None of the research and inquiry items on the MCAS assessment required students to analyze, synthesize, or organize research information.





### ***Speaking and Listening (Criterion B.8)***

Criterion B.8 examines the extent to which the assessment measures students' speaking and listening communication skills. Of the four assessment programs, only Smarter Balanced incorporates listening items and none of the programs assess speaking skills at this time. Because this criterion indicates that programs should assess speaking and listening skills over time and as advances allow, ratings for this criterion were not included in the composite ELA/literacy Content rating.

### ***Depth***

The composite ELA/literacy Depth rating is based on four criteria: Text Quality and Types (B.1), Complexity of Texts (B.2), Cognitive Demand (B.4), and High Quality Items and a Variety of Item Types (B.9). As indicated in Table 9, the assessments for Smarter Balanced (Excellent Match) and ACT Aspire (Good Match) require students to demonstrate the range of thinking skills, including higher-order skills, while the MCAS and PARCC assessments require students to demonstrate less of a range.

***Table 9. Composite ELA/Literacy Depth Ratings***

<b>ACT Aspire</b>	<b>2014 MCAS</b>	<b>PARCC</b>	<b>Smarter Balanced</b>
			

### ***Text Quality and Types (Criterion B.1)***

Criterion B.1 examines the extent to which the assessment requires a balance of high-quality literary and informational texts. The assessment must include the following to fully meet this criterion:

- Approximately two-thirds of the texts at high school are informational and the remainder literary.
- Nearly all passages are high quality (previously published or of publishable quality).
- Nearly all informational passages are expository in structure.



- For grades 6–12, the informational texts are split nearly evenly for literary nonfiction, history/social science, and science/technical.

Table 10 presents how the four programs fared in meeting the requirements for this criterion.

**Table 10. Rating for Text Quality and Types (Criterion B.1)**

ACT Aspire	2014 MCAS	PARCC	Smarter Balanced
G	G	L	E

Smarter Balanced performed exceptionally on this criterion. All passages on this assessment were judged to be high quality, and the assessment included appropriately complex and interesting passages as well as an emphasis on informational rather than expository text. The ACT Aspire and MCAS programs performed well on this criterion. Approximately two-thirds of the texts on the ACT Aspire assessment were informational and nearly all passages were previously published or of publishable quality. Additionally, the majority of informational passages were expository rather than narrative in structure, but the passages were not split approximately evenly for literary nonfiction, history/social science, and science/technical (rather most of the passages were history/social science). The MCAS assessment included appropriate levels of text complexity, but less than two-thirds of the passages were informational and slightly more than half of the informational passages were expository in nature (that is, writing that explains or informs about a specific topic). Additionally, only two of the three writing types (literary nonfiction, history/social science, and science/technical) were addressed rather than having a balance among the three writing types. The PARCC assessment included texts that were of high quality and used open sources, but they were perceived to be overly rigorous. Less than half of the passages on this assessment were informational; however, of the passages that were informational, the majority was expository.

### **Complexity of Texts (Criterion B.2)**

Criterion B.2 examines the extent to which the assessment requires appropriate levels of text complexity, increasing the level each year so that students are ready for the demands of college and career by the end of high school. The documentation for all four programs met the requirements for this criterion; however, due to limitations in the metadata available (see discussion starting on page 48) reviewers were not able to determine the complexity of the actual passages on each form as envisioned by the methodology. Because documentation is not a guarantee of what will appear on actual test forms, ratings for this criterion were not considered when determining the composite Depth rating.




### ***Cognitive Demand (Criterion B.4)***

Criterion B.4 examines the extent to which all students are required to demonstrate a range of higher-order analytical thinking skills in reading and writing based on the depth and complexity of the standards. The assessment must include the following to fully meet this criterion:

- The distribution of cognitive demand on test forms matches the distribution of cognitive demand of the standards as a whole and matches the higher cognitive demand (DOK +3) of the standards.

Table 12 presents how the four programs fared in meeting the requirements for this criterion.

***Table 12. Rating for Cognitive Demand (Criterion B.4)***

ACT Aspire	2014 MCAS	PARCC	Smarter Balanced
			

The ACT Aspire and Smarter Balanced assessments excelled on this criterion. The distribution of cognitive demand for both assessments matched the distribution of cognitive demand of the standards as a whole. Additionally, the percentage of score points associated with DOK levels 3 and 4 approximately matched the percentage of standards at DOK levels 3 and 4. Many items on the MCAS assessment required a lower level of cognitive demand as compared to what was required by the standards, which did not require a high level of strategic or extended thinking. In contrast, the PARCC assessment was seen as overly rigorous because it included a lot of items at the higher DOK levels and few items at the lower DOK levels.<sup>15</sup>

### ***High-Quality Items and a Variety of Item Types (Criterion B.9)***

Criterion B.9 examines the extent to which the assessment uses a variety of item types, including at least one that requires students to generate rather than select a response, and the test items align to the standards and are of high quality. The assessment must include the following to fully meet this criterion:

- At least two item formats are used, including one that requires students to generate rather than select a response.
- All or nearly all operational items reviewed reflect high technical quality, alignment to standards, and high editorial accuracy.

<sup>15</sup> Reviewers in the current study adhered closely to the recommended guidance for rating Criterion B.4. In contrast, Fordham’s reviewers adjusted the rating guidance and they did not rate PARCC lower for including items at higher DOK levels than indicated by the CCSS.

Table 13 presents how the four programs fared in meeting the requirements for this criterion.

**Table 13. Rating for High-Quality Items and a Variety of Item Types (Criterion B.9)**

ACT Aspire	2014 MCAS	PARCC	Smarter Balanced
L	G	E	E

The MCAS, PARCC, and Smarter Balanced programs performed well on this criterion. These three assessments included a variety of item types, with at least one of those types requiring students to generate rather than select a response. Additionally, the items on these three assessment reflected technical quality and editorial accuracy. The PARCC and Smarter Balanced items aligned well to the standards, while the MCAS items were perceived to need improvement in their alignment to the standards. The ACT Aspire assessment included at least two item formats and one of those formats required students to generate a response. However, items on this assessment were perceived to be poorly aligned to the stated grade-level standards and had readability issues due to a lack of specific instructions for responding to the various item types.

### Mathematics

#### Content

The composite mathematics Content rating is based on two criteria: Focus (C.1) and Concepts, Procedures, and Applications (C.2). As shown in Table 14, PARCC and Smarter Balanced received the highest ratings (Excellent Match) and MCAS received a Good Match, indicating these assessments emphasize the most important content of College and Career Ready Standards at the high school level. The early high school ACT Aspire assessment received a Limited rating.

**Table 14. Composite Mathematics Content Ratings**

ACT Aspire	2014 MCAS	PARCC	Smarter Balanced
L	G	E	E

#### Focus (Criterion C.1)

Criterion C.1 examines the extent to which the assessment focuses strongly on the content most needed for success in later mathematics. The assessment must include the following to fully meet this criterion:

- The vast majority (i.e., at least three-quarters at elementary grades, at least two-thirds in middle school grades, and at least half in high school) of score points in

each assessment focuses on the content that is most important for students to master in that grade in order to reach college and career readiness (also called the major work of the grade), and at least 90% of the major work clusters must be assessed by at least one item.

Table 15 presents how the four programs fared in meeting the requirements for this criterion.

**Table 15. Rating for Focus (Criterion C.1)**

ACT Aspire	2014 MCAS	PARCC	Smarter Balanced
L	G	E	E

The PARCC and Smarter Balanced programs performed exceptionally on this criterion while the MCAS program performed well. At least half of the score points on the PARCC and Smarter Balanced assessments focused on widely applicable prerequisites for careers and a wide range of postsecondary studies, which are considered the most important content for students to master. At least half the score points on the MCAS assessment also focused on widely applicable prerequisites for careers and a wide range of postsecondary studies; however, certain standards on the MCAS were assessed multiple times while other standards were not assessed at all. Many of the widely applicable prerequisites assessed on the ACT Aspire assessment were below the high school level and fewer than half the score points were aligned to the high school level widely applicable prerequisites.

**Concepts, Procedures, and Applications (Criterion C.2)**

Criterion C.2 examines the extent to which the test assesses a balance of concepts, procedural skills, and applications. The assessment must include the following to fully meet this criterion:

- On each test form, at least 25% and no more than 50 of score points are allocated to each of the three categories: Mathematical concepts, procedural skill/fluency, and applications.

Table 16 presents how the four programs fared in meeting the requirements for this criterion.

**Table 16. Rating for Concepts, Procedures, and Applications (Criterion C.2)**

ACT Aspire	2014 MCAS	PARCC	Smarter Balanced
W	L	G	G

The PARCC and Smarter Balanced programs performed well on this criterion. Although the distribution of score points devoted to assessing conceptual understanding, procedural skill and fluency, and application was not equally balanced on the PARCC assessment, the items that assessed application were rich in content and practice. The Smarter Balanced assessment included items that assessed conceptual understanding, procedural skill and fluency, and application; however, there was not a balance of the three. Similarly, the MCAS assessment included items that assessed conceptual understanding, procedural skill and fluency, and application, but it did not include a balance. Also, of the MCAS items that assessed conceptual understanding, the complexity of those items was at a very low level and the items that assessed application did not require students to use context to determine meaning or to answer the items. ACT Aspire received a “Weak Match” on this criterion because it had a very low percentage of items that assessed application.

### Depth

The composite mathematics Depth rating is based on three criteria: Connecting Practice to Content (C.3), Cognitive Demand (C.4), and High-Quality Items and a Variety of Item Types (C.5). As can be seen in Table 17, Smarter Balanced (Excellent Match), ACT Aspire, (Good Match) and PARCC (Good Match) fared well on Depth in mathematics. MCAS received a rating of Limited Match.

**Table 17. Composite Mathematics Depth Ratings**

ACT Aspire	2014 MCAS	PARCC	Smarter Balanced
G	L	G	E

### Connecting Practice to Content (Criterion C.3)

Criterion C.3 examines the extent to which the assessment connects mathematical practices to content. The assessment must include the following to fully meet this criterion:

- All or nearly all items that assess mathematical practices also align to one or more content standards.

Table 18 presents how the four programs fared in meeting the requirements for this criterion.

**Table 18. Rating for Connecting Practice to Content (Criterion C.3)**

ACT Aspire	2014 MCAS	PARCC	Smarter Balanced
E	IE	E	E

ACT Aspire, PARCC, and Smarter Balanced excelled on this criterion. All items on these assessments that assessed a mathematical practice also aligned to at least one standard. None of the items on the MCAS assessment specified a mathematical practice and, thus reviewers did not have sufficient evidence (Insufficient Evidence, IE) to provide a rating for this program on this criterion.





### ***Cognitive Demand (Criterion C.4)***

Criterion C.4 examines the extent to which the assessment requires all students to demonstrate a range of higher-order, analytical thinking skills in mathematics based on the depth and complexity of the standards. The assessment must include the following to fully meet this criterion:

- The distribution of cognitive demand on test forms matches the distribution of cognitive demand of the standards as a whole and matches the higher cognitive demand (DOK +3) of the standards.

Table 19 presents how the four programs fared in meeting the requirements for this criterion.

***Table 19. Rating for Cognitive Demand (Criterion C.4)***

<b>ACT Aspire</b>	<b>2014 MCAS</b>	<b>PARCC</b>	<b>Smarter Balanced</b>
			

Smarter Balanced excelled and PARCC performed well on this criterion. The distribution of cognitive demand of the Smarter Balanced assessment matched the distribution of cognitive demand of the standards as a whole and the percentage of score points matched the higher cognitive demand (DOK 3+) of the standards. The distribution of cognitive demand of the items on the PARCC assessment was similar to the distribution of cognitive demand of the standards; however, somewhat more items were needed at the higher DOK levels. For ACT Aspire, the distribution of cognitive demand assessment only partially matched the cognitive demand of the standards. Specifically, reviewers found that both forms included a lower percentage of score points at DOK level 2 than expected by the standards and both forms included a higher percentage of score points at DOK levels 1 and 3 than were expected by the standards. The distribution of cognitive demand for the MCAS assessment was not balanced appropriately; reviewers found there was too much coverage of the lower levels of cognitive demand and not enough coverage of the higher levels of cognitive demand.

### *High-Quality Items and a Variety of Item Types (Criterion C.5)*

Criterion C.5 examines the extent to which the assessment uses a variety of item types, including at least one that requires students to generate rather than select a response, and the test items align to the standards and are of high quality. The assessment must include the following to fully meet this criterion:

- At least two item formats are used, including one that requires students to generate rather than select a response.
- All or nearly all operational items reviewed reflect high technical quality, alignment to standards, and high editorial accuracy.

Table 20 presents how the four programs fared in meeting the requirements for this criterion.

**Table 20. Rating for High-Quality Items and a Variety of Item Types (Criterion C.5)**

ACT Aspire	2014 MCAS	PARCC	Smarter Balanced
L	G	E	G

PARCC excelled while MCAS and Smarter Balanced performed well on this criterion. The PARCC assessment included a variety of item types and one of those types required students to generate as response. Additionally, items on this assessment aligned well to the standards and were technically accurate. The MCAS and Smarter Balanced assessments also included a variety of item types and one of them required students to generate rather than select a response. Some of the MCAS items had technical and editorial issues while a number of the constructed response items had excessive verbiage and required students to have prior knowledge. Similar to the other programs, the ACT Aspire assessment included at least two item types and one of those required students to generate rather than select a response. This program fell short on this criterion because many of its items aligned to off-grade standards and/or had readability issues (e.g., high reading load).

### *Summary of Findings*

#### *ELA/Literacy Content*

Recall the ELA/literacy Content rating is based on Criteria B.3 (Reading), B.5 (Writing), B.6 (Vocabulary and language skills), B.7 (Research and Inquiry), and B.8 (Speaking and Listening). Across the four programs included in this study, Content ratings ranged from Excellent to Weak—the PARCC and Smarter Balanced assessments received an Excellent

Match rating, MCAS received a Limited Match rating, and ACT Aspire received a Weak Match rating. Assessments for all of the programs except ACT Aspire required students to read closely and use evidence from texts (Criterion B.3). The PARCC and Smarter Balanced assessments emphasized writing tasks that required students to engage in close reading and analysis of texts so that students can demonstrate college- and career-readiness abilities (Criterion B.5). The assessments for those same two programs also required students to demonstrate proficiency in the use of language, including vocabulary and conventions (Criterion B.6). All of the assessments except MCAS required students to demonstrate research and inquiry skills by finding, processing, synthesizing, organizing and using information from sources (Criterion B.7). Although the Criteria acknowledge the need to assess speaking and listening skills over time, only the Smarter Balanced assessments currently assess listening skills; none of the programs assessed speaking skills at the time this study was implemented (Criterion B.8).

### ***ELA/Literacy Depth***

The ELA/literacy Depth rating is based on Criteria B.1 (Text Quality and Types), B.2 (Complexity of Texts), B.4 (Cognitive Demand), and B.9 (High-quality Items and Variety of Item Types). The Depth rating was an Excellent Match for Smarter Balanced, Good Match for ACT Aspire, and Limited Match for the other two assessments. For the ELA/literacy assessments, the Smarter Balanced assessments received an Excellent Match rating while the ACT Aspire and MCAS received a Good Match rating, and the PARCC assessment received a Limited Match rating for text quality and balance of types (Criterion B.1). All four programs' assessments required appropriate levels of text complexity and had multiple forms of authentic, previously published texts (Criterion B.2). The ACT Aspire and Smarter Balanced programs had assessments that required students to demonstrate a range of higher-order, analytical thinking skills in reading and writing based on the depth and complexity of college- and career-readiness standards, allowing robust information to be gathered for students with varied levels of achievement (Criterion B.4). All of the programs except ACT Aspire had assessments comprised of high-quality items as defined by the CCSSO criteria that included a variety of item types strategically used to appropriately assess the standards (Criterion B.9).

### ***Mathematics Content***

The mathematics Content rating is based on two criteria, C.1 (Focus) and C.2 (Concepts, Procedures, and Applications). All of the assessments except ACT Aspire focused strongly on the content most needed in high school for later success in mathematics (Criterion C.1). The PARCC and Smarter Balanced assessments measured conceptual understanding,



fluency and procedural skill, and application of mathematics, as indicated in the college- and career-ready standards (Criterion C.2).

### ***Mathematics Depth***

Recall the mathematics Depth rating is based on Criteria B.3 (Close Reading), B.5 (Writing), B.6 (Vocabulary and Language Skills), B.7 (Research and Inquiry), and B.8 (Speaking and Listening). All of the assessments except MCAS included brief questions and longer questions that connected the most important high school mathematical content to mathematical practices (Criterion C.3); insufficient information was provided for the MCAS program to determine the extent to which its assessment connected important content to mathematical practices. The PARCC and Smarter Balanced programs required students to demonstrate a range of higher-order analytical thinking skills, and included questions, tasks, and prompts that measured basic and complex content intended by the college- and career-readiness standards (Criterion C.4). All of the programs except ACT Aspire had assessments with high-quality items as defined by the CCSSO criteria that included a variety of item types strategically used to appropriately assess the standards (Criterion C.5).

### ***Program Responses to Study***

We offered the programs included in this study the opportunity to comment about their participation, including commentary about the results relevant for their assessment and remarks about the test content evaluation methodology. We encouraged each program to include information in their response that might provide background and/or reasoning behind their test design and development as well as any other information that might help interpret this study's results regarding their assessments. The programs' responses are presented in Appendix K.

## Chapter 5: Accessibility Results

As the test content methodology states, the Accessibility review is a “light touch” review, with documentation and only a sample of exemplar items evaluated. The Center’s forthcoming test characteristics methodology, that considers data from administered tests, will support a fuller examination of accessibility. Programs provided extensive documentation for reviewers to consider—ACT Aspire provided 28 documents, MCAS provided 12 documents, PARCC provided 46 documents, and Smarter Balanced provided 35 documents. These documents included information ranging from universal design and accessibility features to user guides and policy documents. Programs highlighted portions of the documentation they felt best addressed each criteria; however, reviewers were not always able to locate the information within the documentation that was relevant to the sub-criterion or, for some instances when the information was located, the reviewers felt it was not sufficiently compelling to fully meet the rating criteria. Programs also provided sets of exemplars for evaluation; however, reviewers were not able to view each exemplar item under all possible accommodated conditions or with all possible accessibility features. Therefore, they did not feel it was appropriate to draw strong conclusions based on their review of only a sample of exemplars. In addition, the Accessibility scoring guidance is very specific and stringent, making it very difficult for any program to receive the highest rating on certain sub-criteria particularly given the aforementioned issues in locating sufficient information to make finely grained judgments. For these reasons, we provide only summary statements for Accessibility in order to prevent over-interpretation of the results.

Program summaries are encapsulated for each program below. The full summary statements can be found in Appendices L-O. For each program, information on strengths is followed by identification of areas for improvement. Other than this general structure, there was no attempt to make the summary statements parallel in content. All programs provided extensive lists of their features. To help understand the breadth of features, we compare the programs on a small sample of accessibility offerings and accommodations (see Table 21).

### **ACT Aspire**

The ACT Aspire summative assessments are administered online or as a paper version, by each state’s choice. The program provides a range of accessibility features and accommodations (e.g., eliminating irrelevant language demand, color contrast, limiting motor load, avoiding extraneous graphics), with similar accessibility features and accommodations offered for the paper-based and online assessments. Documentation includes a rationale for how each feature or accommodation supports valid score interpretations, when each may be used, and

instructions for administration. ACT Aspire demonstrates strong adherence to universal design principles in its development of the assessed content areas. The program presented information about the types of accommodations available (see ACT Aspire link below) and the type of student who might benefit from each based on best practices and research.

It was unclear how the program used information about the types of accommodations available and the type of student who might benefit from each when developing items and assembling forms. Also, the program's implementation of its universal design principles may not have been fully realized during item development and form assembly. For example, reviewers found documentation that indicated the program would provide multiple accommodations but within the documentation provided they were unable to find information about how the program would manage providing multiple accommodations for a single student.

The ACT Aspire Accessibility summary statement can be found in Appendix L. A full list of accessibility features and accommodations offered by ACT Aspire can be found at [http://www.discoveractaspire.org/pdf/2014\\_actaspire\\_Accessibility\\_UserGuide2.0d.pdf](http://www.discoveractaspire.org/pdf/2014_actaspire_Accessibility_UserGuide2.0d.pdf).

### MCAS

The MCAS summative assessments are paper-based. The program offers standard accommodations that change the routine conditions under which a student takes the MCAS (e.g., frequent breaks, unlimited time, magnification, small group) and nonstandard accommodations (modifications) that change a portion of what the test is intended to measure (e.g., read aloud or scribe in ELA, calculator or non-calculator portions of mathematics). These accommodations are provided to students with disabilities as determined by their Individualized Education Plan or 504 Plan and in accordance with the state's participation guidelines. In general, reviewers judged the accommodations and accessibility features offered by MCAS for its summative assessments to be reasonable. MCAS documentation reflected the program's efforts to consider universal design.

There were limited accommodations indicated specifically for ELs. (MCAS provided the *Requirements for the Participation of English Language Learners* document after this study was completed that addressed, at least in part, deficiencies that reviewers found.) Although reviewers judged the accommodations and accessibility features offered by MCAS to be reasonable, they also thought they were limited and did not maintain pace with the field. Currently, the program's use of universal design was perceived to be limited (based on the narrow populations considered and the limited feedback obtained during item development and bias reviews). The program offers a limited scope of accessibility features for some items and certain accommodations appear to

introduce the opportunity for errors because student responses need to be transposed or items had to be skipped. Reviewers did not find a strong connection between research and the accommodations that MCAS made available in the provided documentation. After the study was completed, MCAS clarified that their manuals were written to be accessible and useable by the field; therefore, much of the research studies and policy explanations were not included in them. It is possible these additional documents might have addressed deficiencies that reviewers noted.

The MCAS Accessibility summary statement can be found in Appendix M. A full list of accessibility features and accommodations offered by the 2015-2016 MCAS is available at their website, <http://www.doe.mass.edu/mcas/participation/ell.pdf> and <http://www.doe.mass.edu/mcas/participation/sped.pdf>.

### PARCC

The PARCC summative assessments are administered online and paper-based assessments are offered, as appropriate. The program incorporates accessibility features that are available to all students (e.g., color contrast, eliminate answer choices, highlight tool, pop-up glossary) and offers several test administration considerations for any student (e.g., small group testing, separate location, adaptive and specialized equipment or furniture), as determined by school-based teams. The program also offers a wide range of accommodations for SWDs (e.g., assistive technology, screen reader, Braille note-taker, word prediction external device, extended time) and ELs (e.g., word-to-word dictionary, speech-to-text for mathematics, general directions provided in a student's native language, text-to-speech for the mathematics assessment in Spanish). PARCC was viewed favorably for its sensitivity to the design of item types that reflect individual needs of students with disabilities, and for its strong research base and inclusion of existing research on ELs. Reviewers found the accommodations offered by PARCC to be valid and appropriate based on current research.

Based on the information reviewed during the evaluation, reviewers were unable to locate information about the research needed to determine whether the accessibility features and accommodations that are offered by the program alter the constructs measured in its assessments. Specifically, reviewers noted that clearer documentation may be needed regarding how PARCC administers multiple features simultaneously and the implications of how multiple accessibility features impact student performance. After the workshop, PARCC provided information about how they conduct trials and customer acceptance testing to ensure multiple features and embedded accommodations are properly working that might have addressed deficiencies that reviewers found.

The PARCC Accessibility summary statement can be found in Appendix N. A full list of accessibility features and accommodations offered by PARCC is available on their website, [http://www.parcconline.org/images/Assessments/Accessibility/PARCC\\_Accessibility\\_Features\\_Accommodations\\_Manual\\_v.6\\_01\\_body\\_appendices.pdf](http://www.parcconline.org/images/Assessments/Accessibility/PARCC_Accessibility_Features_Accommodations_Manual_v.6_01_body_appendices.pdf).

### *Smarter Balanced*

The Smarter Balanced summative assessments are administered online as adaptive tests as well as via paper-based versions. The program provides a range of accessibility resources: universal tools, designated supports, and accommodations. Depending on preference, students can select a number of universal tools that are embedded (e.g., digital notepad, highlighter, zoom, English glossary) or non-embedded (e.g., protractor, scratch paper, thesaurus, English glossary) within the assessment.<sup>16</sup> The program also offers a number of designated supports to all students for whom the need has been indicated by an educator or team of educators. The designated supports can be embedded (e.g., color contrast, magnification, translations for the online version, translated glossary) or non-embedded (e.g., color contrast, separate setting, translations for the paper or online versions, translated glossary). For students with documented Individualized Education Plans or 504 Plans, several embedded accommodations are available (i.e., American Sign Language, Braille, closed captioning, and text-to-speech) and several non-embedded accommodations (e.g., abacus, read aloud, scribe, and speech-to-text) are offered. The program has specific guidelines for accessibility for ELs that highlight using clear and accessible language when developing items. Smarter Balanced's use of universal design and evidence-based design were described well. The program also appropriately suggests usability guidance to help educators support determinations of how different accommodations, designated supports and universal tools might interact.

The program's item development procedures incorporated accommodations and accessibility features from conception, which is consistent with the criteria. However, decision making rules were judged to be overly complicated and challenging for educators to apply. For SWDs, certain guidelines were judged to be overly prescriptive when there did not seem to be a reason for such strict guidance. After the workshop, Smarter Balanced highlighted the usability guidance that helps educators support determinations of appropriate accommodations, designated supports and/or universal tools and how they might interact in the *Individual Student Assessment*

---

<sup>16</sup> Embedded supports and accommodations are built into the online test administration and delivery system. Non-embedded supports and accommodations are outside of the online test administration and delivery system.

*Accessibility Profile* documentation. This information may have addressed, at least in part, deficiencies that reviewers noted.

The Smarter Balanced Accessibility summary statement can be found in Appendix O. A full list of accessibility features and accommodations offered by Smarter Balanced is available on their website, [http://www.smarterbalanced.org/wordpress/wp-content/uploads/2014/08/SmarterBalanced\\_Guidelines.pdf](http://www.smarterbalanced.org/wordpress/wp-content/uploads/2014/08/SmarterBalanced_Guidelines.pdf).

### ***Accessibility Feature Comparison***

Within the documentation provided by each program were lists of specifications for each of their accessibility features and accommodations. This documentation is publically available on the programs' websites. Over 60 universal tools, accessibility features, and accommodations were mentioned across the programs. To assist in understanding the Accessibility Review, we compare some of the most common or unique features.

Programs often have many of the same features available, but provide different guidelines on how the features can be accessed and which students are eligible to use them. For example, all programs allow breaks. However, for ACT Aspire and PARCC, breaks usually need to have prior approval by a teacher or administrator. MCAS and Smarter Balanced do not typically require prior approval. For example, when reviewing Smarter Balanced documents, breaks are included as an embedded (for online version) and non-embedded (for paper version) universal tool available to all students without permission.

There were some significant differences in the numbers and types of accessibility features among the programs. Currently, ACT Aspire has the fewest number of accessibility features; they have about 30 for their online assessment and about 35 for their paper-pencil assessment. PARCC and Smarter Balanced have over 50 features listed. However, ACT Aspire indicates they are developing additional features to increase accessibility.

The programs referenced several of the same key documents in their documentation and they sought technical advice from the same industry leaders. Interestingly, one of the state systems that influenced PARCC's accessibility offerings was MCAS. Because PARCC is offered online as well as paper-pencil, the PARCC program is able to offer more accessibility features than MCAS.

The system that ACT Aspire and PARCC use to deliver their tests (Pearson’s TestNav system) might have placed some restrictions on the breadth of accessibility features they could offer. In contrast, the open source system that Smarter Balanced uses to deliver its tests was designed to include more features and tools that increase access. Additionally, Smarter Balanced employs some of the most forward-thinking features for an online test, including streamline and pop-up glossaries. Streamline provides a simplified format that allows items to be displayed directly below the passage or stimuli. For mathematics items pop-up grade appropriate and item specific glossaries are available in 10 languages plus 11 dialects. Both PARCC and Smarter Balanced have expandable passages that allow students to make the passages or stimuli larger. Both allow students to use keyboard shortcuts (e.g., Ctrl+) for personal computers or pinch/zoom for tablets to magnify the screen displays. Clearly, as assessment theory and technological advances permit, programs’ accessibility offerings continue to expand and improve.

Table 21 presents a comparison of the four assessments on key accessibility features and accommodations.

**Table 21. Comparison of Select ELA/Literacy and Mathematics Accessibility Features across the Four Programs**

Feature or Accommodation	ACT Aspire	MCAS	PARCC	Smarter Balanced
American Sign Language of Test	Yes; Writing and Math tests	Yes: Human Signer for ELA and MA items and a CD provided for grade 10 Math.	Yes; ELA/literacy and Math tests	Yes; ELA listening and math tests
Bilingual or Word-to-Word Dictionary <sup>1</sup>	Yes; For any language	Yes; For any language	Yes; For any language	Yes; For any language including grade appropriate pop-up glossaries available in 10 languages plus 11 dialects
Braille and Tactile Graphics	Yes	Yes	Yes	Yes
Breaks	Yes	Yes	Yes	Yes
Closed Captioning	Not applicable	Not applicable	Yes; Transcript for multimedia segments of ELA/literacy test	Yes; For ELA listening items

**Table 21. (Continued)**

Feature or Accommodation	ACT Aspire	MCAS	PARCC	Smarter Balanced
Expandable Passages <sup>2</sup>	No	Not applicable	Yes	Yes
Highlighter	Yes <sup>5</sup>	Yes	Yes	Yes
Speech-to-text <sup>3</sup>	Yes	Yes	Yes	Yes
Streamline <sup>4</sup>	No	Not applicable	No	Yes
Text-to-speech or Read Aloud of items in English	Yes; Writing and Math tests	Yes	Yes, for math test Yes, for ELA tests for SWDs that limits or prevents access to text	Yes, for math test Yes, for ELA tests for SWDs that limits or prevents access to text
Translations of items into other languages	Yes, Writing and Math tests; available as Text-to-Speech Spanish	State law prohibits translations. However, there is one legacy test, the English/Spanish grade 10 math form that has the English items on one page and the Spanish equivalent on the facing page	Spanish math tests is available online and paper versions	Yes; Spanish math tests with stacked translation, and translated mathematics

<sup>1</sup> These do not include definitions, phrases, sentences or pictures.

<sup>2</sup> Passages and stimuli can be expanded so that they take up more of the screen.

<sup>3</sup> Dictated response.

<sup>4</sup> Provides a streamlined, simplified format in which the items are displayed below the stimuli.

<sup>5</sup> When the study was conducted, highlighting was not available in the online version; since then it has been added into ACT's online tools.

All of the programs offer a wide range of accessibility features and accommodations. Some features and universal design tools (e.g., linguistically simplified language for all students) are discussed in research; except for PARCC, these tools were not specifically included in the accessibility/accommodation documentation reviewed. However, this information was found in test development documents for ACT Aspire, MCAS, and Smarter Balanced. Further research is needed by the programs to highlight differences across features and accommodations available for ELs and SWDs and to ensure that the theoretical underpinnings of each assessment include best practices for fairness. Based on discussions with the programs, each plans on conducting research as data become available.



## Chapter 6: Study Challenges and Recommendations

Generally, implementing the methodology for the four programs went smoothly. However, there were a number of challenges that we and the reviewers experienced when implementing this evaluation methodology for the first time. This is not surprising, as it is not unusual to find that not every element works in practice as intended and that fine-tuning is needed. Moreover, any assessment review methodology needs to consider the desire for both a comprehensive and an in-depth review and balance these with the realistic constraints of time and other resources available to conduct the review. In general, the Center balanced efficiency and depth when developing its evaluation methodology.

We discuss some of the challenges of implementing the Center's test content evaluation methodology below and offer recommended revisions for future implementation. Some of our recommendations below might increase efficiencies (e.g., eliminating the step for reviewers to determine individual match scores) but others would necessitate additional reviewer time, which will need to be considered by future implementers.

### *General Challenges*

#### *Testing Program Metadata*

The reliance on testing program metadata for determining which items are rated on certain criteria and/or which items are included in the denominator for subsequent percentages was a challenge, particularly for ELA/literacy. For example, for B.5 (Writing), any item coded by the program to a writing standard was prepopulated into the B.5 rating form. As noted below, this was problematic because not all items that a program aligned to a writing standard required students to actually generate a written response. Consideration and flexibility for different approaches to alignment and metadata coding are needed across the various rating criteria, while still accounting for reviewer burden and rating form usability.

**Recommendation:** We recommend the methodology include a list of detailed metadata requirements for each rating to be made. This will make it clear on which data the ratings are to be based and help identify any particular issues with a vendor's data early in the process.

#### *Redundant Review Process*

The review process consists of multiple steps that involve reviewer discussion and consensus of ratings made during previous steps. We believe two of these steps involve some redundancy and could be collapsed. Specifically, one of the first steps requires individual

reviewers to consider the ratings and comments they provided separately for criteria across both test forms to determine individual match score ratings and then to repeat their consideration of individual match score ratings and comments when determining group match ratings.

***Recommendation:*** We recommend that individual reviewers' tentative match score ratings (which are auto-calculated in Excel spreadsheets) be considered only during group discussion when determining group match score ratings. We found there was no benefit for reviewers to determine individual match ratings because they discussed their rationale when determining final group match score ratings. Eliminating this redundancy could save time in the lengthy review process and help reduce confusion among reviewers.

### ***Small Numbers of Passages Impact Cut-off Percentages***

There are several criteria that are challenging to implement for a single test form, especially Criteria B.1 (Text Quality and Types) and B.2 (Complexity of Texts). Rating category cutoffs are based on percentages that are unstable and do not work well when tests have a small number of texts/passages. For example, if an assessment includes only two informational texts/passages, it can only earn a score of 2 or 0 on Sub-criterion B.1.3.

***Recommendation:*** We recommend that certain criteria be evaluated across multiple forms (as is done for Criterion B.5) rather than evaluated for each form and then aggregated across forms. For instance, for Sub-criterion B.1.3, reviewers would evaluate this criterion across both forms to produce a single rating rather than produce a rating for each test form and then determine a rating across both forms. Evaluating texts/passages across forms will result in a more precise determination of text passage quality and structure. At the same time, rating procedures need to be sensitive to inappropriate variation across forms.

### ***Challenges with Specific Sub-criteria***

#### ***Sub-criterion B.1.2 (Text Quality)***

The purpose of this sub-criterion is to evaluate the quality of texts/passages—whether they are previously published or of publishable quality. However, based on current requirements, there is little to no variability in the ratings that programs receive because essentially all texts/passages included in assessments are previously published. Thus, essentially all assessments will receive high ratings for this criterion.

**Recommendation:** We recommend that future iterations of the methodology consider how to provide a measure of text quality that includes reviewers' judgments in addition to or instead of ratings based on prior publication status of the texts.

### ***Sub-criterion B.2.1 (Justification of Text Based on Data and Qualitative Measures of Complexity)***

Although we excluded Sub-criterion B.2.1 from our study, operationalization of this sub-criterion required that reviewers only confirm that qualitative and quantitative measures of text complexity were used to place the text/passage in the appropriate grade band and level; it did not require reviewers' judgment of text complexity.

**Recommendation:** We recommend that future iterations of the methodology consider having reviewers provide their own judgment of text complexity. We recognize this would require additional resources (e.g., additional training of reviewers, additional time for reviewers to make these judgments), but we believe this would add valuable and confirmatory information.

### ***Sub-criterion B.4.1 and Sub-criterion C.4.1 (Level of Cognitive Demand)***

The methodology calls for the highest rating to be given when the distribution of DOK on a form matches that of the CCSS (a DOK index of .80 or above), and for lower ratings to be assigned as the degree of match declines. However, this does not differentiate between forms where the DOK misalignment is due to the test having too many items of low DOK levels as opposed to the test having too many items of high DOK levels. There was broad agreement among reviewers that tests that include more items with DOK levels 3 and 4 should be rated higher than tests that have more items with DOK levels 1 and 2. This would be consistent with Webb's approach to DOK which is that tests should have at least 50% of the items for a given standard at or above the level required by the standard. Thus tests were not flagged for having too many high DOK items, only for having too few.

**Recommendation:** There are a number of revisions that could alleviate this issue. One possible solution is to include an additional sub-criterion that requires reviewers to evaluate the DOK index of the items compared to the standards to which they are aligned. Another possible solution is to consider the breadth of coverage across multiple forms rather than having reviewers evaluate this criterion separately by form. This would provide additional operationalized data for reviewers to consider when assigning a final B.4 or C.4 rating. Revised guidance could also specify that a lower DOK index threshold would be acceptable if the average DOK of the test exceeds that of the standards. Another approach would be to specify a

*priori* DOK distributions that would be acceptable—for instance, one-third each at levels 1, 2, and 3 and 4.

### ***Sub-criterion B.5.1 (Writing Type) and Sub-criterion B.5.2 (Writing to Sources)***

These criteria are focused on the proportion of prompts that require writing to a source. However, the denominator used to calculate this proportion is the number of items coded to a writing standard regardless of whether the item requires students to actually generate a written response. Thus, depending on how programs code their writing items, some assessments include items that did not require students to generate a written response yet they contributed to the rating for these criteria. Reviewers ultimately used their best judgment and considered only the items that required actual writing when assigning final ratings.

***Recommendation:*** We recommend that future versions of the methodology revise the forms to incorporate only items that require actual writing and include those in the denominator in the calculation for these criteria.

### ***Sub-criterion B.6.3 (Assessing Vocabulary) and Sub-criterion B.6.4 (Assessing Language)***

For Sub-criteria B.6.3 and B.6.4, assessments receive the highest ratings when they report sub-scores for language or vocabulary, or they devote at least 13% of score points to assessing language or vocabulary skills. Based on the current methodology, an assessment can receive the highest score when it reports sub-scores in language or vocabulary even when there are less than 13% of score points devoted to those skills. This occurred for two of the high school assessments—MCAS and PARCC both report language as a sub-score yet both assessments included less than 13% of score points devoted to assessing language skills. Additionally, an assessment that has close to 13% of score points devoted to assessing vocabulary or language yet does not report sub-scores for those skills, should receive a low or the lowest rating. This was the case for the Smarter Balanced assessment—they do not report vocabulary as a sub-score yet the assessment had only slightly less than 13% of score points devoted to these skills. In this instance, reviewers applied their professional judgment to overrule the tentative cut-off. While it is important to provide parents and teachers with direct feedback about students' knowledge of language and vocabulary, assessments should not receive the highest rating if the subscales are based on an unreliably low number of items nor should an assessment receive less than the highest rating if it devotes close to 13% of score points to these skills.

**Recommendation:** We recommend the scoring guidance be revised so that assessments that report language or vocabulary subscales do not receive the highest rating unless they have an adequate number of test items in those areas, which the methodology will need to define.

### **Sub-criterion B.7.1 (Research Skills)**

We experienced two problems with this sub-criterion. First, the definition provided by the methodology is too vague so that reviewers are not clear as to what constitutes items that mirror real-world activities and require students to use research skills to answer them. Second, programs define research various ways, so consistently identifying items that required research skills was a challenge across programs.

**Recommendation:** We recommend that the definition of research item be enhanced to ensure the focus is on the use of two or more discrete sources (e.g., texts/passages) and that research skills are applied in an authentic way. We also suggest there be consideration of including an additional sub-criterion that evaluates the sufficiency of research items (e.g., the percentage of the test devoted to research items). Based on the current methodology, a test that includes a single research item (that covers a very small proportion of the total score points) could receive a high rating.

### **Sub-criterion C.2.1 (Conceptual Understanding, Procedural Skill and Fluency, and Applications)**

We recognize some difficulties might occur when implementing this sub-criterion. Reviewers of the high school assessments were trained to categorize the *predominant* category of the item (conceptual understanding, procedural skill and fluency, and application) and to use the *combined* category sparingly. However, having a combined category as a non-counted option can make it difficult for the distribution of the three item types to match the recommended criteria because of the decreased number of items on which the percentages of each type can be calculated.

**Recommendation:** We recommend that the methodology be revised to allow reviewers to indicate 1, 2, or 3 of the available categories, rather than encouraging them to select just one, which might result in penalizing the assessment if the combined category is selected often. An alternate approach would be to allow reviewers to allocate emphasis across each category (e.g., rate an item as two-thirds procedural and one-third application). If either of these changes were made, the scoring criteria need not be changed—a goal of equal balance of application, procedural skill, and conceptual understanding is still appropriate and can be easily calculated.

### ***Sub-criterion C.3.1 (Connecting Practice and Content)***

This sub-criterion is intended to assess the assessment's measurement of the standards for mathematical practice (SMP); however, it is an inadequate measure of connecting practice to content. The current methodology simply requires that items that are coded to a SMP also be aligned to a content standard (grade level or off-grade). The result is that all assessments either earn the highest rating because the items all code to SMPs or they receive the lowest rating because the items do not code to SMPs. A related concern is that the methodology requires reviewers to merely verify the program's designations of alignment of items to SMPs rather than have the reviewers identify a SMP for the items. Finally, the methodology does not require coverage of the SMPs, allowing a program that assesses only one or two SMPs across all items to receive a high rating.

**Recommendation:** We recommend the methodology be revised to require reviewers to evaluate programs' claims of SMP alignment. This will require additional training and time for review and discussion. We also recommend that the methodology evaluate the extent to which the mathematical practices are adequately covered by the assessment. Again, this could be done using indices of coverage or balance, such as those employed in the Webb alignment methodology. Third, we recommend removing the requirement of content standard coverage from this criterion, as that does not appear to identify any items in practice.

### ***Challenges with Accessibility Review***

Reviewers found it difficult to perform a light touch review when the sub-criteria required them to have a very detailed understanding of the assessment's accessibility features and accommodations to evaluate each of the embedded characteristics. It is unknown how the yet to be finalized test characteristics methodology might be combined with or how it might impact the current test content methodology.

**Recommendation:** If the accessibility review portion of the test content methodology remains a stand-alone review, we recommend the methodology be revised so that the Accessibility review is conducted as an in-depth evaluation to better align with the scoring guidance.

### ***Concluding Commentary***

The four programs that participated in this study made choices about their design and specifications for their assessments (e.g., test length, targeted content). These choices represent operational concerns that are not always in agreement with CCSSO's criteria, yet are reflected in the results obtained from the current study. Further, there are practical concerns

such as testing time and cost that are not included in the criteria, but may be important assessment adoption considerations.

Implementing the test content methodology to evaluate the high school summative assessments for the four programs went smoothly and reviewers felt confident in their ratings. Given we were one of the first organizations to implement this innovative methodology, we were not surprised to find that not every element worked in practice as intended. We have offered a number of recommendations that we hope will enhance the methodology for future implementation.

We are confident that the current study provides valuable information about the quality of the four programs that participated in the study, as evaluated against the CCSSO criteria. We also are hopeful that the recommendations we offer will improve the methodology and provide even more accurate measures to identify high quality assessments.

## Appendix A: CCSSO Criteria for High-Quality Assessments



### CRITERIA for PROCURING and EVALUATING HIGH-QUALITY ASSESSMENTS

States have demonstrated their leadership and commitment to ensuring the success of all students by adopting college- and career-readiness standards. To realize the potential of these standards, states require assessments that match the depth, breadth, and rigor of the standards; accurately measure student progress toward college and career readiness; and provide valid data to inform teaching and learning.

**Assessment of College and Career Readiness.** States have taken different approaches to establishing college- and career-readiness standards and to putting in place high-quality aligned assessments. Many states have adopted the Common Core State Standards (CCSS); some have modified the CCSS to meet their state's context and needs; and others have developed standards independent of the CCSS. To provide assessments that are aligned to these standards, many states are working together through assessment consortia, while others are taking alternative paths for transition. This document is grounded in best practices for assessment development and in the research that defines college and career readiness for English Language Arts(ELA)/literacy and mathematics. Thus, regardless of each state's approach, this document is intended to be a useful resource for any state procuring and/or evaluating assessments aligned to their college- and career-readiness standards.

**Assessment Criteria for States to Consider.** This document provides criteria for states to consider as they develop procurements and evaluate options for high-quality state summative assessments aligned to college- and career-readiness standards. The criteria build on the states' [high-quality summative assessment principles](#) (CCSSO, 2013) which articulate their commitment to high-quality assessments aligned to college and career readiness. To assist states in operationalizing their commitment, this document pays particular attention to not only the criteria states could ask vendors to meet, but also to the evidence states could ask vendors to provide to demonstrate criteria have been – or will be – met. States will, of course, adapt these criteria to reflect their context, standards, and procurement regulations.

**Contents of this Document.** This document begins with an overview of the assessment criteria and continues with a chart containing detailed criteria and sample evidence. These criteria do not cover every area that a state would have to address in a procurement or evaluation process. Instead, they focus on the critical characteristics that should be met by high-quality assessments aligned to college- and career-readiness standards. A more comprehensive source for the development and validation of assessments is the *Standards for Educational and Psychological Testing* (AERA, APA, and NCME, 1999). The assessment criteria and evidence discussed herein were developed by referencing the *Standards for Educational and Psychological Testing* and several other key sources listed in the bibliography. Additional state-specific criteria at the end of the document highlight a few of the most important additional issues that states may wish to consider in a procurement or evaluation process.

**Notes about Evidence and Terminology.** This document is intended to support states in selecting assessments that meet a high bar for quality. Thus, the document suggests the evidence that states will need to review in order to make informed judgments on vendors' claims about the quality of their proposed assessments. Of course, vendors may propose assessments that are yet to be developed, assessments in development, and/or existing assessments. In designing procurement or evaluation procedures, states may therefore find it helpful to design the process for awarding "points" so as neither to reward existing (but poor quality) tests just because they have data available, nor to reward well-intentioned conceptual designs that are not executable. To support this goal, vendors should be asked to provide the most rigorous level of evidence they have available, consistent with the stage of assessment development they are in. The types of evidence that vendors should be expected to provide at different stages of development are described below:



- For assessments to be newly created, the most rigorous level of evidence will include the vendor’s descriptions of their established and proven processes; data from similar assessments; proposed test blueprints and other specifications (e.g., test design documents, test specifications, item specifications, scoring specifications); exemplar test items, passages, and forms; proposed studies, reports, and technical documentation to be created during assessment development and operation; and the processes for responding to such data. In addition, the vendor’s prior experience, expertise, and letters of recommendation should be included.
- For assessments that are currently in development, the most rigorous level of evidence will depend on the stage of assessment development. Evidence should include test blueprints and other specifications (e.g., test design documents, test specifications, item specifications, scoring specifications), and exemplar test items, passages, and forms. In addition, evidence should include as much of the data described below regarding pre-existing assessments as is available. Where such evidence is not available, vendors should provide descriptions of their established and proven processes; data from similar assessments, proposed studies, reports, and technical documentation to be created during assessment development and operation; and the process for responding to such data. In addition, the vendor’s prior experience, expertise, and letters of recommendation should be included.
- For pre-existing assessments, the most rigorous level of evidence will include comprehensive validity evidence; test blueprints and other specifications (e.g., test design documents, test specifications, item specifications, scoring specifications); annual technical reports; results of studies on scaling, equating, and reporting; and exemplar test items, passages, and forms.

Additionally, regardless of the stage of test development, states may find it helpful to put in place best practice quality assurance and other processes so that states can monitor quality throughout development and administration, and periodically evaluate evidence to ensure criteria are being met.

Finally, a note about terminology. In this document, the term “assessments” generally refers to the entire suite of summative assessments a state would procure – that is, tests of ELA/literacy and mathematics in each grade assessed. In sections specifically about ELA/literacy or mathematics, however, the term refers to the set of summative assessments in that content area. The terms “assessment” and “test” are often used interchangeably when discussing a single grade level/content area. Throughout the document, the term “tasks” refers to extended-response, open-ended test items; “test items” refers to the stimuli used to elicit a response through, for example, multiple-choice or constructed-response items as well as tasks; and “forms” are systematic collections of test items and tasks that comprise the testing experience for a particular student in a grade/content area.

## Overview of Assessment Criteria

### A. Meet Overall Assessment Goals and Ensure Technical Quality

- A.1 Indicating progress toward college and career readiness
- A.2 Ensuring that assessments are valid for required and intended purposes
- A.3 Ensuring that assessments are reliable
- A.4 Ensuring that assessments are designed and implemented to yield valid and consistent test score interpretations within and across years
- A.5 Providing accessibility to *all* students, including English learners and students with disabilities
- A.6 Ensuring transparency of test design and expectations
- A.7 Meeting all requirements for data privacy and ownership

### B. Align to Standards – English Language Arts/Literacy

- B.1 Assessing student reading and writing achievement in both ELA and literacy
- B.2 Focusing on complexity of texts
- B.3 Requiring students to read closely and use evidence from texts
- B.4 Requiring a range of cognitive demand
- B.5 Assessing writing
- B.6 Emphasizing vocabulary and language skills
- B.7 Assessing research and inquiry
- B.8 Assessing speaking and listening
- B.9 Ensuring high-quality items and a variety of item types

### C. Align to Standards – Mathematics

- C.1 Focusing strongly on the content most needed for success in later mathematics
- C.2 Assessing a balance of concepts, procedures, and applications
- C.3 Connecting practice to content
- C.4 Requiring a range of cognitive demand
- C.5 Ensuring high-quality items and a variety of item types

### D. Yield Valuable Reports on Student Progress and Performance

- D.1 Focusing on student achievement and progress to readiness
- D.2 Providing timely data that inform instruction

### E. Adhere to Best Practices in Test Administration

- E.1 Maintaining necessary standardization and ensuring test security

### F. State Specific Criteria (as desired)

*Sample criteria might include*

- Requiring involvement of the state’s K-12 educators and institutions of higher education
- Procuring a system of aligned assessments, including diagnostic and interim assessments
- Ensuring interoperability of computer-administered items

### Assessment Criteria and Evidence

#### A. Meet Overall Assessment Goals and Ensure Technical Quality<sup>\*</sup>

Criteria	Evidence
<b>A.1 Indicating progress toward college and career readiness:</b> Scores and performance levels on assessments are mapped to determinations of college and career readiness at the high school level and for other grades to being on track to college and career readiness by the time of high school graduation.	<ul style="list-style-type: none"> <li>• A description is provided of the process for developing performance level descriptors and setting performance standards (i.e., “cut scores”), including             <ul style="list-style-type: none"> <li>○ Appropriate involvement of higher education and career/technical experts in determining the score at which there is a high probability that a student is college and career ready;</li> <li>○ External evidence used to inform the setting of performance standards and a rationale for why certain forms of evidence are included and others are not (e.g., student performance on current state assessments, NAEP, TIMSS, PISA, ASVAB, ACT, SAT, results from Smarter Balanced and PARCC, relevant data on post-secondary performance, remediation, and workforce readiness);</li> <li>○ Evidence and a rationale that the method(s) for including external benchmarks are valid for the intended purposes; and</li> <li>○ Standard setting studies, the resulting performance level descriptors and performance standards, and the specific data on which they are based (when available).</li> </ul> </li> <li>• A description is provided of the intended studies that will be conducted to evaluate the validity of performance standards over time.</li> </ul>
<b>A.2 Ensuring that assessments are valid for required and intended purposes:</b> Assessments produce data, including student achievement data and student growth data required under Title I of the Elementary and Secondary Education Act (ESEA) and ESEA Flexibility, that can be used to validly inform the following: <ul style="list-style-type: none"> <li>• School effectiveness and improvement;</li> <li>• Individual principal and teacher effectiveness for purposes of evaluation and identification of professional development and support needs;</li> <li>• Individual student gains and performance; and</li> <li>• Other purposes defined by the state.</li> </ul>	<ul style="list-style-type: none"> <li>• A well-articulated validity evaluation based on an interpretive argument (e.g., Kane, 2006) is provided that includes, at a minimum             <ul style="list-style-type: none"> <li>○ Evidence of the validity of using results from the assessments for the three primary purposes, as well as any additional purposes required by the state (specify sources of data).</li> <li>○ Evidence that scoring and reporting structures are consistent with structures of the state’s standards (specify sources of data).</li> <li>○ Evidence that total test and relevant sub-scores are related to external variables as expected (e.g., other measures of the construct). To the extent possible, include evidence that the items are “instructionally sensitive,” that is, that item performance is more related to the quality of instruction than to out-of-school factors such as demographic variables.</li> <li>○ Evidence that the assessments lead to the intended outcomes (i.e., meet the intended purposes) and minimize unintended negative consequences. Consequential evidence</li> </ul> </li> </ul>

<sup>\*</sup> The term “technical quality” here refers to the qualities necessary to ensure that scoring and generalization inferences based on test scores are valid both within and across years. This document prioritizes certain aspects of technical quality, but as noted in the introduction, readers should also refer to other sources, primarily *The Standards for Educational and Psychological Testing*.

Criteria	Evidence
	<p>should flow from a well-articulated theory of action about how the assessments are intended to work and be integrated with the larger accountability system.</p> <ul style="list-style-type: none"> <li>○ The set of content standards against which the assessments are designed is provided. If these standards are the state’s standards, evidence is provided that the content of the assessments reflects the standards, including the cognitive demand of the standards. If they are not the state’s standards, evidence is provided of the extent of alignment with the state’s standards.</li> <li>○ Evidence is provided to ensure the content validity of test forms and the usefulness of score reports (e.g., test blueprints demonstrate the learning progressions reflected in the standards, and experts in the content and progression toward readiness are significantly involved in the development process).</li> </ul>
<b>A.3 Ensuring that assessments are reliable:</b> Assessments minimize error that may distort interpretations of results, estimate the magnitude of error, and inform users of its magnitude.	<ul style="list-style-type: none"> <li>• Evidence is provided of the reliability of assessment scores, based on the state’s student population and reported subpopulations (specify sources of data).</li> <li>• Evidence is provided that the scores are reliable for the intended purposes for essentially all students, as indicated by the standard error of measurement across the score continuum (i.e., conditional standard error).</li> <li>• Evidence is provided of the precision of the assessments at cut scores, and consistency of student level classification (specify sources of data).</li> <li>• Evidence is provided of generalizability for all relevant sources, such as variability of groups, internal consistency of item responses, variability among schools, consistency from form to form of the test, and inter-rater consistency in scoring (specify sources of data).</li> </ul>
<b>A.4 Ensuring that assessments are designed and implemented to yield valid and consistent test score interpretations within and across years:</b>	
<ul style="list-style-type: none"> <li>• <b>Assessment forms</b> yield consistent score meanings over time, forms within year, student groups, and delivery mechanisms (e.g., paper, computer, including multiple computer platforms).</li> </ul>	<ul style="list-style-type: none"> <li>• A description is provided of the process used to ensure comparability of assessments and assessment results across groups and time.</li> <li>• Evidence is provided of valid and reliable linking procedures to ensure that the scores derived from the assessments are comparable within year across various test “forms” and across time.</li> <li>• Evidence is provided that the linking design and results are valid for test scores across the achievement continuum.</li> </ul>
<ul style="list-style-type: none"> <li>• <b>Score scales</b> used facilitate accurate and meaningful inferences about test performance.</li> </ul>	<ul style="list-style-type: none"> <li>• Evidence is provided that the procedures used to transform raw scores to scale scores is coherent with the test design and the intended claims, including the types of Item Response Theory (IRT) calibration and scaling methods (if used) and other methods for facilitating meaningful score interpretations over tests and time.</li> <li>• Evidence is provided that the assessments are designed and scaled to ensure the primary</li> </ul>

Criteria	Evidence
	<p>interpretations of the assessment can be fulfilled. For example, if the assessments are used as data sources for growth or value-added models for accountability purposes, evidence should be provided that the scaling and design features would support such uses, such as ensuring appropriate amounts of measurement information throughout the scale, as appropriate.</p> <ul style="list-style-type: none"> <li>Evidence is provided, where a vertical or other score scale is used, that the scaling design and procedures lead to valid and reliable score interpretations over the full length of the scale proposed; and evidence is provided that the scale is able to maintain these properties over time (or a description of the proposed procedures is provided).</li> </ul>
<p><b>A.5 Providing accessibility to all students, including English learners and students with disabilities:</b></p> <ul style="list-style-type: none"> <li><b>Following the principles of universal design:</b> The assessments are developed in accordance with the principles of universal design and sound testing practice, so that the testing interface, whether paper- or technology-based, does not impede student performance.</li> </ul>	<ul style="list-style-type: none"> <li>A description is provided of the item development process used to reduce construct irrelevance (e.g., eliminating unnecessary clutter in graphics, reducing construct-irrelevant reading load as much as possible), including <ul style="list-style-type: none"> <li>The <i>test item</i> development process to remove potential challenges due to factors such as disability, ethnicity, culture, geographic location, socioeconomic condition, or gender; and</li> <li><i>Test form</i> development specifications that ensure that assessments are clear and comprehensible for all students.</li> </ul> </li> <li>Evidence is provided, including exemplar tests (paper and pencil forms or screen shots) illustrating principles of universal design.</li> </ul>
<ul style="list-style-type: none"> <li><b>Offering appropriate accommodations and modifications:</b> Allowable accommodations and modifications that maintain the constructs being assessed are offered where feasible and appropriate, and consider the access needs (e.g., cognitive, processing, sensory, physical, language) of the vast majority of students.</li> </ul>	<ul style="list-style-type: none"> <li>A description is provided of the accessibility features that will be available, consistent with state policy (e.g., magnification, audio representation of graphic elements, linguistic simplification, text-to-speech, speech-to-text, Braille).</li> <li>A description is provided of access to translations and definitions, consistent with state policy.</li> <li>A description is provided of the construct validity of the available accessibility features with a plan that ensures that the scores of students who have accommodations or modifications that do not maintain the construct being assessed are not combined with those of the bulk of students when computing or reporting scores.</li> </ul>
<ul style="list-style-type: none"> <li>Assessments produce valid and reliable scores for <b>English learners</b>.</li> </ul>	<ul style="list-style-type: none"> <li>Evidence is provided that test items and accessibility features permit English learners to demonstrate their knowledge and abilities and do not contain features that unnecessarily prevent them from accessing the content of the item. Evidence should address: presentation, response, setting, and timing and scheduling (specify sources of data).</li> </ul>
<ul style="list-style-type: none"> <li>Assessments produce valid and reliable scores for <b>students with disabilities</b>.</li> </ul>	<ul style="list-style-type: none"> <li>Evidence is provided that test items and accessibility features permit students with disabilities to demonstrate their knowledge and abilities and do not contain features that</li> </ul>

Criteria	Evidence
	<p>unnecessarily prevent them from accessing the content of the item. Evidence should address: presentation, response, setting, and timing and scheduling (specify sources of data).</p>
<p><b>A.6 Ensuring transparency of test design and expectations:</b> Assessment design documents (e.g., item and test specifications) and sample test questions are made publicly available so that all stakeholders understand the purposes, expectations, and uses of the college- and career-ready assessments.</p>	<ul style="list-style-type: none"> <li>Evidence is provided, including test blueprints, showing the range of state standards covered, reporting categories, and percentage of assessment items and score points by reporting category.</li> <li>Evidence is provided, including a release plan, showing the extent to which a representative sample of items will be released on a regular basis (e.g., annually) across every grade level and content area.</li> <li>Sample items with annotations and answer rationales are provided.</li> <li>Scoring rubrics for constructed-response items with sample responses are provided for each level of the rubric.</li> <li>Item development specifications are provided.</li> <li>Additional information is provided to the state to demonstrate the overall quality of the assessment design, including <ul style="list-style-type: none"> <li>Estimated testing time by grade level and content area;</li> <li>Number of forms available by grade level and content area;</li> <li>Plan for what percentage of items will be refreshed and how frequently;</li> <li>Specifications for the various levels of cognitive demand and how each is to be represented by grade level and content area; and</li> <li>For ELA/Literacy, data from text complexity analyses.</li> </ul> </li> </ul>
<p><b>A.7 Meeting all requirements for data privacy and ownership:</b> All assessments must meet federal and state requirements for student privacy, and all data is owned exclusively by the state.</p>	<ul style="list-style-type: none"> <li>An assurance is provided of student privacy protection, reflecting compliance with all applicable federal and state laws and requirements.</li> <li>An assurance is provided of state ownership of all data, reflecting knowledge of state laws and requirements.</li> <li>An assurance is provided that the state will receive all underlying data, in a timely and useable fashion, so it can do further analysis as desired, including, for example, achievement, verification, forensic, and security analyses.</li> <li>A description is provided for how data will be managed securely, including, for example, as data is transferred between vendors and the state.</li> </ul>

**B. Align to Standards – English Language Arts/Literacy**

Criteria	Evidence
<b>B.1 Assessing student reading and writing achievement in both ELA and literacy:</b> The assessments are English language arts and literacy tests that are based on an aligned balance of high-quality literary and informational texts.	<ul style="list-style-type: none"> <li>• Test blueprints and other specifications as well as exemplar literary and informational passages are provided for each grade level, demonstrating the expectations below are met.</li> <li>• Texts are balanced across literary and informational text types and across genres, with more informational than literary texts used as the assessments move up in the grade bands, as the state's standards require. <i>For example, for common core aligned assessments, goals include</i> <ul style="list-style-type: none"> <li>○ <i>In grades 3-8, approximately half of the texts are literature and half are informational;</i></li> <li>○ <i>In high school, because comprehension of complex informational texts is crucial for readiness, texts are approximately one-third literature and two-thirds informational; and</i></li> <li>○ <i>In all grades, informational texts are primarily expository rather than narrative in structure, and in grades 6-12, informational texts are approximately one-third each literary nonfiction, history/social science, and science/technical.</i></li> </ul> </li> <li>• Texts and other stimuli (e.g., audio, visual, graphic) are previously published or of publishable quality. They are content-rich, exhibit exceptional craft and thought, and/or provide useful information.</li> <li>• History/social studies and science/technical texts, specifically, reflect the quality of writing that is produced by authorities in the particular academic discipline.</li> </ul>
<b>B.2 Focusing on complexity of texts:</b> The assessments require appropriate levels of text complexity; they raise the bar for text complexity each year so students are ready for the demands of college- and career-level reading no later than the end of high school. Multiple forms of authentic, previously published texts are assessed, including written, audio, visual, and graphic, as technology and assessment constraints permit.	<ul style="list-style-type: none"> <li>• Text complexity measurements, exemplar literary and informational passages for each grade level, and other evidence (e.g., data, tools, procedures) are provided to demonstrate the expectations below are met.</li> <li>• At each grade, reading texts have sufficient complexity, and the average complexity of texts increases grade-by-grade, meeting college- and career-ready levels by the end of high school.</li> <li>• A rationale and evidence are provided for how text complexity is quantitatively and qualitatively measured and used to place each text at the appropriate grade level. <i>For example, for common core aligned assessments, goals include</i> <ul style="list-style-type: none"> <li>○ <i>Texts are placed in a grade band using at least one research-based quantitative measure;</i></li> <li>○ <i>Texts are placed at a grade level using a qualitative analysis measure, reflecting the expert judgment of educators; and</i></li> <li>○ <i>Most of the texts are placed within the grade band indicated by the quantitative</i></li> </ul> </li> </ul>

Criteria	Evidence
<b>B.3 Requiring students to read closely and use evidence from texts:</b> Reading assessments consist of test questions or tasks, as appropriate, that demand that students read carefully and deeply and use specific evidence from increasingly complex texts to obtain and defend correct responses.	<p style="text-align: center;"><i>analysis, with exceptions usually found in high school literary texts.</i></p> <ul style="list-style-type: none"> <li>• Test blueprints and other specifications as well as exemplar test items are provided for each grade level, demonstrating the expectations below are met.</li> <li>• All reading questions are text-dependent and             <ul style="list-style-type: none"> <li>○ Arise from and require close reading and analysis of text;</li> <li>○ Focus on the central ideas and important particulars of the text, rather than on superficial or peripheral concepts; and</li> <li>○ Assess the depth and specific requirements delineated in the standards at each grade level (i.e., the concepts, topics, and texts specifically named in the grade-level standards).</li> </ul> </li> <li>• Many reading questions require students to directly provide textual evidence in support of their responses. <i>For example, for common core aligned assessments, goals include</i> <ul style="list-style-type: none"> <li>○ <i>A majority of reading score points is devoted to questions that ask students to directly provide textual evidence in support of their responses (e.g., constructed-response and/or two-part evidence-based selected-response item formats).</i></li> </ul> </li> </ul>
<b>B.4 Requiring a range of cognitive demand:</b> The assessments require all students to demonstrate a range of higher-order, analytical thinking skills in reading and writing based on the depth and complexity of college- and career-ready standards, allowing robust information to be gathered for students with varied levels of achievement.	<ul style="list-style-type: none"> <li>• Test blueprints and other specifications are provided to demonstrate that the distribution of cognitive demand for each grade level and content area is sufficient to assess the depth and complexity of the state's standards, as evidenced by use of a generic taxonomy (e.g., Webb's Depth of Knowledge) or, preferably, classifications specific to the discipline and drawn from the requirements of the standards themselves and item response modes, such as             <ul style="list-style-type: none"> <li>○ The complexity of the text on which an item is based;</li> <li>○ The range of textual evidence an item requires (how many parts of text[s] students must locate and use to respond to the item correctly);</li> <li>○ The level of inference required; and</li> <li>○ The mode of student response (e.g., selected-response, constructed-response).</li> </ul> </li> <li>• A rationale is provided justifying the distribution of cognitive demand for each grade level and content area.</li> <li>• Exemplar test items for each grade level are provided, illustrating each level of cognitive demand, and accompanied by a description of the process used to determine an item's cognitive level.</li> </ul>
<b>B.5 Assessing writing:</b> Assessments emphasize writing tasks that require students to engage in close reading and analysis of texts so that students can demonstrate college- and career-ready abilities.	<ul style="list-style-type: none"> <li>• Test blueprints and other specifications as well as exemplar test items for each grade level are provided, demonstrating the expectations below are met.</li> <li>• Writing tasks reflect the types of writing that will prepare students for the work required in college and the workplace, balancing expository, persuasive/argument, and narrative writing, as state standards require. At higher grade levels, the balance shifts toward more exposition and argument.</li> </ul>

Criteria	Evidence
	<p><i>For example, for common core aligned assessments, goals include</i></p> <ul style="list-style-type: none"> <li>○ <i>Taking all forms of the test together, writing tasks are approximately one-third each exposition, argument, and narrative (some tasks may represent blended structures), with the balance shifting toward more exposition and argument at the higher grade levels.</i></li> </ul> <ul style="list-style-type: none"> <li>• <b>Tasks (including narrative tasks) require students to confront text or other stimuli directly, to draw on textual evidence, and to support valid inferences from text or stimuli.</b></li> </ul>
<p><b>B.6 Emphasizing vocabulary and language skills:</b> The assessments require students to demonstrate proficiency in the use of language, including vocabulary and conventions.</p>	<ul style="list-style-type: none"> <li>• Test blueprints and other specifications as well as exemplar test items for each grade level are provided, demonstrating the expectations below are met.</li> <li>• Vocabulary items reflect requirements for college and career readiness, including <ul style="list-style-type: none"> <li>○ Focusing on general academic (tier 2) words;</li> <li>○ Asking students to use context to determine meaning; and</li> <li>○ Assessing words that are important to the central ideas of the text.</li> </ul> </li> <li>• Language is assessed within writing assessments as part of the scoring rubric, or it is assessed with test items that specifically address language skills. Language assessments reflect requirements for college and career readiness by <ul style="list-style-type: none"> <li>○ Mirroring real-world activities (e.g., actual editing or revision, actual writing); and</li> <li>○ Focusing on common student errors and those conventions most important for readiness.</li> </ul> </li> <li>• Assessments place sufficient emphasis on vocabulary and language skills (i.e., a significant percentage of the score points is devoted to these skills).</li> </ul>
<p><b>B.7 Assessing research and inquiry:</b> The assessments require students to demonstrate research and inquiry skills, demonstrated by the ability to find, process, synthesize, organize, and use information from sources.</p>	<ul style="list-style-type: none"> <li>• Test blueprints and other specifications as well as exemplar test items for each grade level are provided, demonstrating the expectations below are met.</li> <li>• Test items assessing research and inquiry mirror real world activities and require students to analyze, synthesize, organize, and use information from sources. <i>For example, for common core aligned assessments, goals include</i> <ul style="list-style-type: none"> <li>○ <i>Research tasks require writing to sources, including analyzing, selecting, and organizing evidence from more than one source, and often from sources in diverse formats; and</i></li> <li>○ <i>When assessment constraints permit, real or simulated research tasks comprise a significant percentage of score points when all forms of the reading and writing test are considered together.</i></li> </ul> </li> </ul>
<p><b>B.8 Assessing speaking and listening:</b> Over time, and as assessment advances allow, the assessments measure the speaking and listening communication skills students need for college and career readiness.</p>	<ul style="list-style-type: none"> <li>• Over time, and as assessment advances allow, the speaking and listening skills required for college and career readiness are assessed. <i>For example, for common core aligned assessments, test items assessing speaking</i> <ul style="list-style-type: none"> <li>○ <i>Assess students' ability to express well-supported ideas clearly and to probe others' ideas; and</i></li> </ul> </li> </ul>

Criteria	Evidence
	<ul style="list-style-type: none"> <li>○ <i>Include items that measure students' ability to marshal evidence from research and orally present findings in a performance task.</i></li> </ul> <p><i>For example, for common core aligned assessments, test items assessing listening</i></p> <ul style="list-style-type: none"> <li>○ <i>Are based on texts and other stimuli that meet the criteria for complexity, range, and quality outlined in criteria B.1 and B.2 above; and</i></li> <li>○ <i>Permit the evaluation of active listening skills (e.g., taking notes on main ideas, elaborating on remarks of others).</i></li> </ul>
<p><b>B.9 Ensuring high-quality items and a variety of item types:</b> High-quality items and a variety of types are strategically used to appropriately assess the standard(s).</p>	<ul style="list-style-type: none"> <li>• Specifications are provided to demonstrate that the distribution of item types for each grade level and content area is sufficient to strategically assess the depth and complexity of the standards being addressed. Item types may include, for example, selected-response, two-part evidence-based selected-response, short and extended constructed-response, technology-enhanced, and performance tasks.</li> <li>• To support claims of quality, the following are provided: <ul style="list-style-type: none"> <li>○ Exemplar items for each item type used in each grade band;</li> <li>○ Rationales for the use of the specific item types;</li> <li>○ Specifications showing the proportion of item types on a form;</li> <li>○ For constructed response and performance tasks, a scoring plan (e.g., machine-scored, hand-scored, by whom, how trained), scoring rubrics, and sample student work to confirm the validity of the scoring process; and</li> <li>○ A description of the process used for ensuring the technical quality, alignment to standards, and editorial accuracy of the items.</li> </ul> </li> </ul>

C. Align to Standards – Mathematics

Criteria	Evidence
<p><b>C.1 Focusing strongly on the content most needed for success in later mathematics:</b> The assessments help educators keep students on track to readiness by focusing strongly on the content most needed in each grade or course for later mathematics.</p>	<ul style="list-style-type: none"> <li>• Test blueprints and other specifications are provided, demonstrating that the vast majority of score points in each assessment focuses on the content that is most important for students to master in that grade band in order to reach college and career readiness. For each grade band, this content consists of               <ul style="list-style-type: none"> <li>○ Elementary grades – number and operations;</li> <li>○ Middle school – ratio, proportional relationships, pre-algebra, and algebra; and</li> <li>○ High school – prerequisites for careers and a wide range of postsecondary studies, particularly algebra, functions, and modeling applications.</li> </ul> </li> </ul> <p><i>For example, for common core aligned assessments, goals include</i></p> <ul style="list-style-type: none"> <li>○ <i>In elementary grades, at least three-quarters of the points in each grade align exclusively to the major work of the grade;</i></li> <li>○ <i>In middle school grades, at least two-thirds of the points in each grade align exclusively</i></li> </ul>

Criteria	Evidence
	<p><i>to the major work of the grade; and</i></p> <ul style="list-style-type: none"> <li>○ <i>In high school, at least half of the points in each course align exclusively to prerequisites for careers and a wide range of postsecondary studies.</i></li> </ul> <ul style="list-style-type: none"> <li>• The assessment design reflects the state’s standards and reflects a coherent progression of mathematics content from grade to grade and course to course.</li> </ul>
<p><b>C.2 Assessing a balance of concepts, procedures, and applications:</b> The assessments measure conceptual understanding, fluency and procedural skill, and application of mathematics, as set out in college- and career-ready standards.</p>	<ul style="list-style-type: none"> <li>• Test blueprints and other specifications as well as exemplar test items for each grade level are provided, demonstrating the expectations below are met.</li> <li>• The distribution of score points reflects a balance of mathematical concepts, procedures/fluency, and applications, as the state’s standards require.</li> </ul> <p><i>For example, for common core aligned assessments, at least one-quarter of the points come from each of the following categories:</i></p> <ul style="list-style-type: none"> <li>○ <i>Conceptual understanding problems in which students to respond to well-designed conceptual problems;</i></li> <li>○ <i>Procedural skill and fluency problems (e.g., purely procedural problems, some requiring use of efficient algorithms, and others inviting opportunistic strategies); and</i></li> <li>○ <i>Application problems (e.g., in elementary and middle grades, solving grade-appropriate word problems reflecting growing complexity across the grades; in high school, rich application problems requiring students to demonstrate college and career readiness).</i></li> </ul> <ul style="list-style-type: none"> <li>• All students, whether high performing or low performing, are required to respond to items within the categories of conceptual understanding, procedural skill and fluency, and applications, so they have the opportunity to show what they know and can do.</li> </ul>
<p><b>C.3 Connecting practice to content:</b> The assessments include brief questions and also longer questions that connect the most important mathematical content of the grade or course to mathematical practices, for example, modeling and making mathematical arguments.</p>	<ul style="list-style-type: none"> <li>• Test blueprints and other specifications as well as exemplar test items for each grade level are provided, demonstrating the expectations below are met.</li> <li>• Assessments for each grade and course meaningfully connect mathematical practices and processes with mathematical content (especially with the most important mathematical content at each grade), as required by the state’s standards.</li> <li>• Explanatory materials (citing test blueprints and other specifications) describe the connection for each grade or course between content and mathematical practices and processes.</li> </ul> <p><i>For example, for common core aligned assessments, goals include</i></p> <ul style="list-style-type: none"> <li>○ <i>Every test item that assesses mathematical practices is also aligned to one or more content standards (most often within the major work of the grade); and</i></li> <li>○ <i>Through the grades, test items reflect growing sophistication of mathematical practices with appropriate expectations at each grade level.</i></li> </ul>

Criteria	Evidence
<p><b>C.4 Requiring a range of cognitive demand:</b> The assessments require all students to demonstrate a range of higher-order, analytical thinking skills in reading and writing based on the depth and complexity of college- and career-ready standards, allowing robust information to be gathered for students with varied levels of achievement. Assessments include questions, tasks, and prompts about the basic content of the grade or course as well as questions that reflect the complex challenge of college- and career-ready standards.</p>	<ul style="list-style-type: none"> <li>• Test blueprints and other specifications are provided to demonstrate that the distribution of cognitive demand for each grade level is sufficient to assess the depth and complexity of the state’s standards, as evidenced by use of a generic taxonomy (e.g., Webb’s Depth of Knowledge) or, preferably, classifications specific to the discipline and drawn from mathematical factors, such as               <ul style="list-style-type: none"> <li>○ Mathematical topic coverage in the task (single topic vs. two topics vs. three topics vs. four or more topics);</li> <li>○ Nature of reasoning (none, simple, moderate, complex);</li> <li>○ Nature of computation (none, simple numeric, complex numeric or simple symbolic, complex symbolic);</li> <li>○ Nature of application (none, routine word problem, non-routine or less well-posed word problem, fuller coverage of the modeling cycle); and</li> <li>○ Cognitive actions (knowing or remembering, executing, understanding, investigating, or proving).</li> </ul> </li> <li>• A rationale is provided justifying the distribution of cognitive demand for each grade level and content area.</li> <li>• Exemplar test items for each grade level are provided, illustrating each level of cognitive demand, and accompanied by a description of the process used to determine an item’s cognitive level.</li> </ul>
<p><b>C.5 Ensuring high-quality items and a variety of item types:</b> High-quality items and a variety of item types are strategically used to appropriately assess the standard(s).</p>	<ul style="list-style-type: none"> <li>• Specifications are provided to demonstrate that the distribution of item types for each grade level and content area is sufficient to strategically assess the depth and complexity of the standards being addressed. Item types may include selected-response, short and extended constructed-response, technology-enhanced, and multi-step problems.</li> <li>• To support claims of quality the following are provided:               <ul style="list-style-type: none"> <li>○ The list and distribution of the types of work students will be asked to produce (e.g., facts, computation, diagrams, models, explanations);</li> <li>○ Exemplar items for each item type used in each grade band;</li> <li>○ Rationales for the use of the specific item types;</li> <li>○ Specifications showing the proportion of item types on a form;</li> <li>○ For constructed response items, a scoring plan (e.g., machine-scored, hand-scored, by whom, how trained), scoring rubrics, and sample student work to confirm the validity of the scoring process; and</li> <li>○ A description of the process used for ensuring the technical quality, alignment to standards, and editorial accuracy of the items.</li> </ul> </li> </ul>

#### D. Yield Valuable Reports on Student Progress and Performance

Criteria	Evidence
<p><b>D.1 Focusing on student achievement and progress to readiness:</b> Score reports illustrate a student’s progress on the continuum toward college and career readiness, grade by grade, and course by course. Reports stress the most important content, skills, and processes, and how the assessment focuses on them, to show whether or not students are on track to readiness.</p>	<ul style="list-style-type: none"> <li>• A list of reports is provided, and for each report, a sample that shows, at a minimum               <ul style="list-style-type: none"> <li>○ Scores and sub-scores that will be reported with emphasis on the most important content, skills, and processes for each grade or course;</li> <li>○ Explanations of results that are instructionally valuable and easily understood by essentially all audiences;</li> <li>○ Results expressed in terms of performance standards (i.e., proficiency “cut scores”), not just scale scores or percentiles; and</li> <li>○ Progress on the continuum toward college and career readiness, which can be expressed by whether a student has sufficiently mastered the current grade or course content and is therefore prepared for the next level.</li> </ul> </li> </ul> <p>(Note: Not all reporting information need be numerical; for example, actual student work on a released item could be presented, along with the rubric for the item and a discussion of common errors.)</p> <ul style="list-style-type: none"> <li>• The reporting structure can be supported by the assessment design, as demonstrated by evidence, including data confirming that test blueprints include a sufficient number of items for each reporting category, so that scores and sub-scores lead to the intended interpretations and minimize the possibility of misinterpretation.</li> </ul>
<p><b>D.2 Providing timely data that inform instruction:</b> Reports are instructionally valuable, easy to understand by all audiences, and delivered in time to provide useful, actionable data to students, parents, and teachers.</p>	<ul style="list-style-type: none"> <li>• A timeline and other evidence are provided to show when assessment results will be available for each report.</li> <li>• A description is provided of the process and technology that will be used to issue reports in as timely a manner as possible.</li> <li>• Evidence, including results of user testing, is provided to demonstrate the utility of the reports for each intended audience.</li> </ul>

#### E. Adhere to Best Practices in Test Administration

Criteria	Evidence
<p><b>E.1 Maintaining necessary standardization and ensuring test security:</b> In order to ensure the validity, fairness, and integrity of state test results, the assessment systems maintain the security of the items and tests as well as the answer documents and related ancillary materials that result from test administrations.</p>	<ul style="list-style-type: none"> <li>• A comprehensive security plan is provided with auditable policies and procedures for test development, administration, score reporting, data management, and detection of irregularities consistent with NCES and CCSSO recommendations for, at a minimum               <ul style="list-style-type: none"> <li>○ Training for all personnel – both test developers and administrators;</li> <li>○ Secure management of assessments and assessment data, so that no individual gains access to unauthorized information;</li> <li>○ Test administration and environment; and</li> <li>○ Methods used to detect testing irregularities before, during, and after testing, and steps</li> </ul> </li> </ul>



Criteria	Evidence
	<p>to address them.</p> <ul style="list-style-type: none"> <li>A description is provided of how security safeguards have been tested and validated for computer-based tests and for paper-and-pencil tests, as relevant.</li> </ul>

**F. State Specific Criteria** (as desired)

*It is likely that states will supplement the above criteria with criteria specific to their needs. These might, for example, include*

- **Requiring involvement of the state’s K-12 educators, institutions of higher education, and career/technical experts** in the design, development, and/or scoring of the assessments;
- **Procuring a system of aligned assessments, including diagnostic and interim assessments** designed to target and improve instruction as well as measure progress and performance; and
- **Ensuring interoperability of computer-administered items** consistent in all ways with the specifications laid out in the *Assessment Interoperability Framework* (2012) developed by the Common Education Data Standards (CEDS) project, so that tests and items owned by the state can be easily ported from one technology platform to another.

## Bibliography

- American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) Joint Committee on Standards for Educational and Psychological Testing. (1999). *Standards for educational and psychological testing*. Washington DC: AERA. Retrieved March 21, 2014, from <http://www.apa.org/science/programs/testing/standards.aspx>.
- Chingos, M.M. (2013). *Standardized testing and the common core standards: You get what you pay for?* Washington, DC: Brown Center on Education Policy at Brookings. Retrieved March 21, 2014, from <http://www.brookings.edu/research/reports/2013/10/30-standardized-testing-and-the-common-core-chingos>.
- CCSSO. (2013). *States' commitment to high-quality assessments aligned to college- and career-readiness*. Washington, DC: Author. Retrieved March 21, 2014, from <http://www.ccsso.org/Documents/2013/CCSSO%20Assessment%20Quality%20Principles%202010-1-13%20FINAL.pdf>.
- CCSSO and the Association of Test Publishers (ATP). (2010). *Operational best practices for statewide large-scale assessment programs*. Washington, DC: Author. Retrieved March 21, 2014, from [http://www.ccsso.org/resources/publications/operational\\_best\\_practices\\_for\\_statewide\\_large-scale\\_assessment\\_programs.html](http://www.ccsso.org/resources/publications/operational_best_practices_for_statewide_large-scale_assessment_programs.html).
- Common Education Data Standards (CEDs). (2012). *Assessment interoperability framework*. Washington, DC: Author. Retrieved March 21, 2014, from <https://ceds.ed.gov/aif.aspx>.
- Darling-Hammond, L., & Adamson, F. (2013). *Developing assessments of deeper learning: The costs and benefits of using tests that help students learn*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education. Retrieved March 21, 2014, from <http://edpolicy.stanford.edu/publications/pubs/745>.
- Darling-Hammond, L., Herman, J., Pellegrino, J., et al. (2013). *Criteria for high-quality assessment*. Stanford, CA: Stanford Center for Opportunity. Retrieved March 21, 2014, from <https://edpolicy.stanford.edu/publications/pubs/847>.
- Kane, M.T. (2006). Validation. In R.L. Brennan (Ed.), *Educational measurement* (4<sup>th</sup> ed., pp. 17-64). Washington, DC: The National Council on Measurement in Education and the American Council on Education.
- Michigan Department of Education. (2013). *Report on options for assessments aligned with the common core state standards, submitted to the Michigan Legislature*. Lansing, MI: Michigan Department of Education. Retrieved March 21, 2014, from [http://www.michigan.gov/documents/mde/Common\\_Core\\_Assessment\\_Option\\_Report\\_441322\\_7.pdf](http://www.michigan.gov/documents/mde/Common_Core_Assessment_Option_Report_441322_7.pdf).
- Student Achievement Partners. (2013). *Assessment evaluation tool for CCSS alignment in ELA/literacy grades 3-12*. NYC: Author. Retrieved March 21, 2014, from <http://www.achievethecore.org/page/298/aet-ela-literacy-grades-3-12>.
- Student Achievement Partners. (2013). *Assessment evaluation tool for CCSS alignment in mathematics grades K-HS*. NYC: Author. Retrieved March 21, 2014, from <http://www.achievethecore.org/page/297/aet-mathematics-grades-k-hs>.
- U.S. Department of Education. (2010). *Overview information: Race to the top fund, notice inviting applications for new awards for fiscal year (FY) 2010*. Washington, DC: Author. Retrieved March 21, 2014, from <http://www2.ed.gov/programs/racetothetop-assessment/resources.html>.
- U.S. Department of Education. (2013). *Race to the top assessment program, technical review process, April 2013*. Washington, DC: Author. Retrieved March 21, 2014, from <http://www2.ed.gov/programs/racetothetop-assessment/technical-review-process.pdf>.
- U.S. Department of Education, Institute of Education Sciences National Center for Education Statistics. (2013). *Testing integrity symposium: Issues and recommendations for best practice*. Washington, DC: Author. Retrieved March 21, 2014, from <http://nces.ed.gov/pubs2013/2013454.pdf>.
- U.S. Department of Education, Office of Elementary and Secondary Education. (2007). *Standards and assessments peer review guidance: Information and examples for meeting requirements of the No Child Left Behind Act of 2001*. Washington, DC: Author. Retrieved March 21, 2014, from <http://www2.ed.gov/policy/elsec/guid/saaprguidance.pdf>.

## Appendix B: ELA/Literacy Scoring Template

### List of Criteria and Sub-Criteria

Sub-Criteria	Type
<b>Criterion B.1 (Depth)</b>	
B.1.1 Informational and literary text balance	Outcome
B.1.2 Text quality	Outcome
B.1.3 Type of informational texts	Outcome
B.1.4 Specification of informational and literary balance	Generalizability
B.1.5 Specification of quality of texts	Generalizability
B.1.6 Specification of type of informational texts	Generalizability
<b>Criterion B.2 (Depth)</b>	
B.2.1 Justification of texts based on data and qualitative measures of complexity	Outcome
B.2.2 Procedures and rationale for how text complexity is measured	Generalizability
B.2.3 Documentation specifies target text complexity	Generalizability
<b>Criterion B.3 (Content)</b>	
B.3.1 Close reading	Outcome
B.3.2 Central ideas and important particulars	Outcome
B.3.3 Questions text dependent and assess depth	Outcome
B.3.4 Questions require direct textual evidence	Outcome
B.3.5 Specification on text-dependency	Generalizability
B.3.6 Specification on proportion of scores devoted to textual evidence	Generalizability
<b>Criterion B.4 (Depth)</b>	
B.4.1 Level of cognitive demand	Outcome
B.4.2 Procedures for evaluating cognitive demand	Generalizability
<b>Criterion B.5 (Content)</b>	
B.5.1 Percentages of writing type	Outcome
B.5.2 Percentages of prompts requiring writing to sources	Outcome
B.5.3 Specification of distribution of writing tasks/types	Generalizability
B.5.4 Specifications require confrontation with texts/stimuli directly	Generalizability
<b>Criterion B.6 (Content)</b>	
B.6.1 Vocabulary using tier 2 words, require use of text, and important to central ideas	Outcome
B.6.2 Mirror real-world activities, focus on common errors, and emphasize conventions	Outcome
B.6.3 Percentage of score points devoted to assessing vocabulary	Outcome
B.6.4 Percentage of score points devoted to assessing language	Outcome
B.6.5 Specifications for vocabulary for college and career readiness	Generalizability
B.6.6 Specifications of points for vocabulary	Generalizability
B.6.7 Specification of distribution of vocabulary	Generalizability
B.6.8 Specifications place sufficient emphasis on vocabulary	Generalizability
<b>Criterion B.7 (Content)</b>	
B.7.1 %age of research skills items requiring analysis, synthesis, &/or organization of info	Outcome
B.7.2 Significance of research	Generalizability
B.7.3 Specifications on real/simulated research tasks	Generalizability
<b>Criterion B.8 (Content)</b>	
B.8.1 Items based on listening skills	Outcome
B.8.2 Items based on speaking skills	Outcome
B.8.3 Specifications on listening skills	Generalizability
B.8.4 Specification on speaking skills	Generalizability
<b>Criterion B.9 (Depth)</b>	
B.9.1 Kinds of formats used on operational forms	Outcome
B.9.2 Quality of items	Outcome
B.9.3 Specifications of item type	Generalizability
B.9.4 Specifications of quality	Generalizability

**B.1 Assessing student reading and writing achievement in both ELA and literacy:** The assessments are English language arts and literacy tests that are based on an aligned balance of high-quality literary and informational texts.

	Type	Evidence Descriptors	Location of Evidence	Scoring Guidelines	Tentative Cut-Offs
B.1.1	Outcome	<p>Texts are balanced across literary and informational text types and across genres, with more informational than literary texts used as the assessments move up in the grade bands.</p> <p>Goals include;</p> <ul style="list-style-type: none"> <li>In grades 3-8, approximately half of the texts are literature and half are informational.</li> <li>In high school, because comprehension of complex informational texts is crucial for readiness, texts are approximately one-third literature and two-thirds informational.</li> </ul>	<p>Evidence: Test forms, meta-data</p> <p>Coding Sheets:</p> <ul style="list-style-type: none"> <li>Is the passage informational or literary?</li> </ul> <p>Metrics Auto-Calculated:</p> <ul style="list-style-type: none"> <li>Percent of passages informational.</li> <li>Percent of passages literary.</li> </ul>	<p>Calculate the percentage of informational texts vs. literary texts on the reading and writing assessments (not language skills assessments). Assign a score and provide notes under Comments (for each form):</p> <p>Assign a score for <u>grades 3-8</u>:</p> <p><b>2 – Meets:</b> Approximately half of the texts are informational.  <b>1 – Partially Meets:</b> At least one-third of the texts are informational.  <b>0 – Does Not Meet:</b> Less than one-third or nearly all of the texts are informational.</p> <p>Assign a score for <u>high school</u>:</p> <p><b>2 –Meets:</b> Approximately two-thirds of the texts are informational.  <b>1 – Partially Meets:</b> Less than approximately two-thirds are informational.  <b>0 – Does Not Meet:</b> Less than half or nearly all of the texts are informational.</p> <p>Note: Because the percentage of informational text should increase as students move up through the grades, it is also appropriate for the percentages of informational texts in grades 6-8 to be closer to the high school guidelines as students prepare for reading more informational texts in high school.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p>For grades 3 - 8:  <b>2 – Meets:</b> 45-55%  <b>1 – Partially Meets:</b> 33-44% or 56-84%.  <b>0 – Does Not Meet:</b> 0-32% or 85-100%.</p> <p>For <u>high school</u> grades:  <b>2 –Meets:</b> 60-72%.  <b>1 – Partially Meets:</b> 40-59% or 73-90%.  <b>0 – Does Not Meet:</b> 0-39% or 91-100%</p>

<b>B.1 Assessing student reading and writing achievement in both ELA and literacy:</b> The assessments are English language arts and literacy tests that are based on an aligned balance of high-quality literary and informational texts.					
	Type	Evidence Descriptors	Location of Evidence	Scoring Guidelines	Tentative Cut-Offs
<b>B.1.2</b>	Outcome	Texts and other stimuli (e.g., audio, visual, graphic) are previously published or of publishable quality. They are content-rich, exhibit exceptional craft and thought, and/or provide useful information.	Evidence: Test forms, meta-data  Coding Sheet <ul style="list-style-type: none"> <li>Is the passage is previously published (Y/N)</li> <li>If not previously published, is the passage of publishable quality? (Y/N)</li> </ul> Metrics Auto-Calculated: <ul style="list-style-type: none"> <li>Number/% of previously published passages</li> <li>Number/% of passages of publishable quality</li> </ul>	If the writing test does not employ passages, the rating will be based on reading passages only. Calculate the percentage of passages that meet the quality criteria. Assign a score and provide notes under Comments (for each form):  <b>2 –Meets:</b> Nearly all passages are high quality (previously published or of publishable quality). <b>1 – Partially Meets:</b> The large majority of passages (i.e. three-quarters or more) are high quality (previously published or of publishable quality). <b>0 – Does Not Meet:</b> Less than the large majority of passages are high quality (previously published or of publishable quality).  Definition: Publishable quality texts are content-rich, exhibit exceptional craft and thought, and/or provide useful information.  Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.	<b>2 – Meets:</b> 90-100% <b>1 – Partially Meets:</b> 75-89% <b>0 – Does Not Meet:</b> 0-74%
<b>B.1.3</b>	Outcome	In all grades, informational texts are primarily expository rather than narrative in structure, and in grades 6-12, informational texts are approximately one-third each literary nonfiction, history/social science, and science/technical.	Evidence: Test forms and meta-data  Coding Sheet: <ul style="list-style-type: none"> <li>If the passage is informational, is the structure primarily narrative or expository? (Narrative/Expository)               <ul style="list-style-type: none"> <li>If the passage is informational, which discipline best describes the passage content (Literary Nonfiction; History/Literary Nonfiction; Science and Technical/Literary Nonfiction; History/Science and Technical;</li> </ul> </li> </ul>	For informational texts at ALL grades, calculate the number of passages that are primarily expository in structure. For informational texts at grades 6-12, calculate the balance of literary nonfiction, history/social science, and science/technical texts. Assign a score and provide notes under Comments (for each form):  <b>2- Meets:</b> Nearly all informational passages are expository in structure AND for grades 6-12, the informational texts are split nearly evenly for literary nonfiction, history/social science, and science/technical. <b>1 – Partially Meets:</b> The large majority of informational passages (i.e., three-quarter) are expository in structure AND/OR for grades 6-12, the informational texts address only two of the three disciplines mentioned above. <b>0 – Does Not Meet:</b> Less than the large majority of informational passages (i.e., less than three-quarters) are expository in structure AND/OR for grades 6-12, the informational texts address only one of the three disciplines mentioned above.	<b>2 – Meets:</b> 90-100% are expository AND for grades 6-12, the informational texts are split nearly evenly for literary nonfiction, history/social science, and science/technical <b>1 – Partially Meets:</b> 75-89% are expository AND/OR for grades 6-12, the informational texts address only two of the three disciplines mentioned above. <b>0 – Does Not Meet:</b> 0-74% are expository AND/OR for grades 6-12, the informational texts address only one of the three disciplines mentioned above.

**B.1 Assessing student reading and writing achievement in both ELA and literacy:** The assessments are English language arts and literacy tests that are based on an aligned balance of high-quality literary and informational texts.

	Type	Evidence Descriptors	Location of Evidence	Scoring Guidelines	Tentative Cut-Offs
			<p>History/Science and Technical/Literary Nonfiction Informational Passages)</p> <p>Metrics Auto-Calculated:</p> <ul style="list-style-type: none"> <li>• Number and percent of informational passages with a narrative structure</li> <li>• Number and percent of informational passages with an expository structure</li> <li>• Number and percent of history informational passages</li> <li>• Number and percent of science/technical informational passages</li> <li>• Number and percent of literary nonfiction informational passages</li> <li>• Number and percent of History/Literary nonfiction informational passages</li> <li>• Number and percent of science and technical/literary nonfiction informational passages</li> <li>• Number and percent of history/science and technical informational passages</li> <li>• Number and percent of history/science and technical/literary nonfiction informational passages</li> </ul>		

<b>B.1 Assessing student reading and writing achievement in both ELA and literacy:</b> The assessments are English language arts and literacy tests that are based on an aligned balance of high-quality literary and informational texts.					
	<b>Type</b>	<b>Evidence Descriptors</b>	<b>Location of Evidence</b>	<b>Scoring Guidelines</b>	<b>Tentative Cut-Offs</b>
<b>B.1.4</b>	Generalizability	<p>Test blueprints and/or other specifications specify for each grade level the proportions of each text type and genre each student should be administered.</p> <p>The test blueprints distribution of emphasis of text types follows the CCSSO <i>Criteria</i>. Goals include:</p> <ul style="list-style-type: none"> <li>• Texts are balanced across literary and informational text types and across genres, with more informational than literary texts used as the assessments move up in the grade bands.</li> <li>▪ In grades 3-8, approximately half of the texts are literature and half are informational;</li> <li>▪ In high school, texts are approximately one-third literature and two-thirds informational;</li> <li>▪ In all grades, informational texts are primarily expository rather than narrative in structure, and in grades 6-12, informational texts are approximately one-third each literary nonfiction, history/social science, and science/technical.</li> </ul>	Evidence: Test blueprints and/or other documents identified by the program.	<p>Rate the extent to which the documentation represents the distributions of the type of passages. Assign a score and provide notes under Comments:</p> <p>Assign a score for <u>grades 3-8</u>:</p> <p><b>2 – Meets:</b> Specifications indicate that approximately half of the texts should be informational.  <b>1 – Partially Meets:</b> Specifications indicate that at least one-third of the texts should be informational.  <b>0 – Does Not Meet:</b> Specifications indicate that less than one-third or nearly all of the texts should be informational.</p> <p>Assign a score for <u>high school</u>:</p> <p><b>2 –Meets:</b> Specifications indicate that approximately two-thirds of the texts should be informational.  <b>1 – Partially Meets:</b> Specifications indicate that less than approximately two-thirds should be informational.  <b>0 – Does Not Meet:</b> Specifications indicate that less than half or nearly all of the texts should be informational.</p> <p>Note: Because the percentage of informational text should increase as students move up through the grades, it is also appropriate for the percentages of informational texts in grades 6-8 to be closer to the high school guidelines as students prepare for reading more informational texts in high school.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p><u>For grades 3-8:</u>  <b>2 –Meets:</b> 45-55%  <b>1 – Partially Meets:</b> 33-44% or 56-84%  <b>0 – Does Not Meet:</b> 0-32% or 85-100%</p> <p><u>For high school:</u>  <b>2 –Meets:</b> 60-72%  <b>1 – Partially Meets:</b> 40-59% or 72-90%  <b>0 – Does Not Meet:</b> 0-39% or 91-100%</p>

<b>B.1 Assessing student reading and writing achievement in both ELA and literacy:</b> The assessments are English language arts and literacy tests that are based on an aligned balance of high-quality literary and informational texts.					
	<b>Type</b>	<b>Evidence Descriptors</b>	<b>Location of Evidence</b>	<b>Scoring Guidelines</b>	<b>Tentative Cut-Offs</b>
<b>B.1.5</b>	Generalizability	<p>As part of the construct definition, the quality of texts is defined. The program’s definitions are consistent with the <i>CCSSO Criteria</i>:</p> <ul style="list-style-type: none"> <li>• Texts and other stimuli (e.g., audio, visual, graphic) are previously published or of publishable quality.</li> <li>• They are content-rich, exhibit exceptional craft and thought, and/or provide useful information.</li> <li>• History/social studies and science/technical texts, specifically, reflect the quality of writing that is produced by authorities in the particular academic discipline.</li> </ul>	Evidence: Test blueprints and/or other documents identified by the program.	<p>Rate the extent to which the construct definition and the quality of the texts are specified in the documents. Assign a score and provide notes under Comments:</p> <p><b>2 –Meets:</b> Specifications indicate that nearly all passages should be of high quality (previously published or of publishable quality).</p> <p><b>1 – Partially Meets:</b> Specifications indicate that a large majority of passages (i.e., three-quarters or more) should be of high quality (previously published or of publishable quality).</p> <p><b>0 – Does Not Meet:</b> Specifications indicate that less than the large majority of passages should be of high quality (previously published or of publishable quality).</p> <p>If the writing test will not use passages, the rating will be based on reading passages only.</p> <p>Definition: Publishable quality texts are content-rich, exhibit exceptional craft and thought, and/or provide useful information.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p><b>2 – Meets:</b> 90-100%</p> <p><b>1 – Partially Meets:</b> 75-89%</p> <p><b>0 – Does Not Meet:</b> 0-74%</p>
<b>B.1.6</b>	Generalizability	In all grades, informational texts are primarily expository rather than narrative in structure, and in grades 6-12, informational texts are approximately one-third each literary nonfiction, history/social science, and science/technical.	Evidence: Test blueprints and/or other documents identified by the program	<p>Rate the extent to which the documents require that informational texts be expository in structure and for grades 6-12, the distributions of text by disciplines is addressed. Assign a score and provide notes under Comments:</p> <p><b>2- Meets:</b> Documentation outlines that for all grades, informational passages should be primarily expository in structure AND for grades 6-12, the informational texts are split nearly evenly for literary nonfiction, history/social science, and science/technical.</p> <p><b>1 – Partially Meets:</b> Documentation outlines EITHER that informational passages are primarily expository in structure OR that for grades 6-12, the informational texts should be split nearly evenly for literary nonfiction, history/social science, and science/technical.</p> <p><b>0 – Does Not Meet:</b> Documentation does not outline requirements for informational texts that are expository in structure nor are there requirements for including a balance of literary nonfiction, history/social science, and science/technical texts.</p>	<p><b>2 – Meets:</b> 90-100% are expository AND for grades 6-12, the informational texts are split nearly evenly for literary nonfiction, history/social science, and science/technical.</p> <p><b>1 – Partially Meets:</b> 75-89% are expository OR for grades 6-12, the informational texts are split nearly evenly for the three disciplines mentioned above.</p> <p><b>0 – Does Not Meet:</b> 0-74% are expository AND for grades 6-12, the informational texts are not balanced in the three disciplines mentioned above.</p>



<b>B.2 Focusing on complexity of texts:</b> The assessments require appropriate levels of text complexity; they raise the bar for text complexity each year so students are ready for the demands of college- and career-level reading no later than the end of high school. Multiple forms of authentic, previously published texts are assessed, including written, audio, visual, and graphic, as technology and assessment constraints permit.					
	Type	Evidence Descriptors	Location of Evidence	Scoring Guidelines	Tentative Cut-Offs
B.2.1	Outcome	<p>Text complexity is quantitatively and qualitatively measured and used to place each text at the appropriate grade level.</p> <p>Goals include:</p> <ul style="list-style-type: none"> <li>• Texts are placed in a grade <b>band</b> using at least one research-based quantitative measure;</li> <li>• Texts are placed at a grade <b>level</b> using a qualitative analysis measure, reflecting the expert judgment of educators; and</li> <li>• Most of the texts are placed within the grade <b>band</b> indicated by the quantitative analysis, with exceptions usually found in high school literary texts</li> </ul> <p>See Common Core State Standards Appendix A regarding text complexity.</p>	<p>Evidence: Test forms, meta-data</p> <p>Coding Sheet</p> <ul style="list-style-type: none"> <li>• Is there evidence of both quantitative and qualitative analysis? (Y/N)</li> <li>• Is the passage placed in appropriate grade band based on quantitative data? (Y/N or N/A)</li> <li>• Is the passage placed in appropriate grade level based on qualitative analysis? (Y/N)</li> </ul> <p>Metrics Auto-Calculated:</p> <ul style="list-style-type: none"> <li>• Number and percent of texts placed in correct grade band based on quantitative data</li> <li>• Number and percent of texts placed in correct grade level based on qualitative data</li> <li>• Number and percent of texts placed in correct grade band based on quantitative data AND in correct grade level based on qualitative analysis</li> </ul>	<p>Determine the percentage of passages placed at a grade <u>band</u> that is justified by quantitative data and a grade <u>level</u> justified by qualitative measures. Assign a score and provide notes under Comments (for each form):</p> <p><b>2 – Meets:</b> All or nearly all passages have been placed at a grade band and grade level justified by complexity data.  <b>1 – Partially Meets:</b> A large majority of passages (i.e., three quarters or more) have been placed at a grade band and grade level justified by complexity data.  <b>0 – Does Not Meet:</b> Less than a large majority of passages have been placed at a grade band justified by complexity data</p> <p>“Complexity data” refers to results from both quantitative and qualitative measures.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p><b>2 – Meets:</b> 90-100%  <b>1 – Partially Meets:</b> 75-89%  <b>0 – Does Not Meet:</b> 0-74%</p>

**B.2 Focusing on complexity of texts:** The assessments require appropriate levels of text complexity; they raise the bar for text complexity each year so students are ready for the demands of college- and career-level reading no later than the end of high school. Multiple forms of authentic, previously published texts are assessed, including written, audio, visual, and graphic, as technology and assessment constraints permit.

	Type	Evidence Descriptors	Location of Evidence	Scoring Guidelines	Tentative Cut-Offs
B.2.2	Generalizability	<p>Procedures and a rationale are provided for how text complexity is quantitatively and qualitatively measured, and a procedure defines how to place each text at the appropriate grade level.</p> <p>Goals include:</p> <ul style="list-style-type: none"> <li>▪ Texts are placed in a grade band using at least one research-based quantitative measure;</li> <li>▪ Texts are placed at a grade level using a qualitative analysis measure, reflecting the expert judgment of educators; and</li> <li>▪ Most of the texts are placed within the grade band indicated by the quantitative analysis, with exceptions usually found in high school literary texts.</li> </ul>	Evidence: Test blueprints and/or other documents identified by the program.	<p>Evaluate whether the documentation indicates the percentage of passages placed at a grade <u>band</u> that is justified by quantitative data and a grade <u>level</u> justified by qualitative measures. Assign a rating and provide notes under Comments:</p> <p>2- Meets: – The documentation clearly explains how quantitative data is used to determine grade band placement AND texts are then placed at the grade level recommended by qualitative review. Text complexity rating process results in nearly all passages being placed at a grade band and grade level justified by complexity data.*</p> <p>1 – Partially Meets: The documentation explains only how either quantitative data is used to determine grade band OR qualitative data is used to determine grade level placement. Text complexity rating process results in the large majority (i.e., three quarters or more) passages being placed at a grade band and grade level justified by complexity data.*</p> <p>0 – Does Not Meet: The documentation does not explain the relationship of quantitative data to grade band or qualitative data to grade level placement. Text complexity rating process results in less than the large majority of passages being placed at a grade band and grade level justified by complexity data.*</p> <p>*In rare instances, qualitative analysis may overrule quantitative data in grade band placement. These specific places are poetry and drama (across all grades), and literature (in high school only).</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p><b>2 – Meets:</b> 90-100%</p> <p><b>1 – Partially Meets:</b> 75-89%</p> <p><b>0 – Does Not Meet:</b> 0-74%</p>

**B.2 Focusing on complexity of texts:** The assessments require appropriate levels of text complexity; they raise the bar for text complexity each year so students are ready for the demands of college- and career-level reading no later than the end of high school. Multiple forms of authentic, previously published texts are assessed, including written, audio, visual, and graphic, as technology and assessment constraints permit.

	Type	Evidence Descriptors	Location of Evidence	Scoring Guidelines	Tentative Cut-Offs
<b>B.2.3</b>	Generalizability	Documentation specifies that the average target complexity of texts increases grade-by-grade, meeting college- and career-ready levels by the end of high school.	Evidence: Test blueprints and/or other documents identified by the program.	<p>Rate the extent to which the documentation specifies that the average target complexity of texts increases grade-by-grade, meeting college- and career-ready levels by the end of high school. Assign a rating and provide notes under Comments:</p> <p><b>2 –Meets:</b> Documentation outlines that text complexity increases by grade <u>level</u> across all years of the assessment program, meeting CCR levels by end of high school.</p> <p><b>1 – Partially Meets:</b> Documentation outlines that text complexity increases by grade <u>band</u> across all years of the assessment program, meeting CCR levels by end of high school.</p> <p><b>0 – Does Not Meet:</b> Documentation does not outline a requirement for increasing text complexity as students progress through the grades to ensure they meet CCR levels by end of high school.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p><b>2 – Meets:</b> details progression by grade level</p> <p><b>1 – Partially Meets:</b> details progression by grade band only</p> <p><b>0 – Does Not Meet:</b> does not include details about increasing text complexity</p>

<b>B.3 Requiring students to read closely and use evidence from texts:</b> Reading assessments consist of test questions or tasks, as appropriate, that demand that students read carefully and deeply and use specific evidence from increasingly complex texts to obtain and defend correct responses.					
	<b>Type</b>	<b>Evidence Descriptors</b>	<b>Location of Evidence</b>	<b>Scoring Guidance</b>	<b>Tentative Cut-Scores</b>
<b>B.3.1</b>	Outcome	All reading questions are text-dependent and arise from and require close reading and analysis of text.	Evidence: Test forms, meta-data  Specific metadata from assessment program: <ul style="list-style-type: none"> <li>▪ Assigned CCSS alignment (and secondary alignment(s), if any)</li> </ul> Point value of item Coding Sheets: <ul style="list-style-type: none"> <li>▪ Is the item aligned to the specifics of the standard? (Y/N)</li> <li>▪ Does item require close reading and analysis? (Y/N)</li> <li>▪ Does item focus on central ideas and important particulars? (Y/N)</li> <li>▪ Does the item require direct use of textual evidence? (Y/N)</li> </ul>	Determine the percentage of items that require close reading and analysis of text rather than skimming, recall, or simple recognition of paraphrased text. Assign a rating and provide notes under Comments (for each form):  <b>2 – Meets:</b> Nearly all items require close reading and analysis of text. <b>1 – Partially Meets:</b> The large majority of items (i.e., three-quarters or more) require close reading and analysis of text. <b>0 – Does Not Meet:</b> Less than a large majority of the items require close reading and analysis of text.  Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.	<b>2 – Meets:</b> 90-100% <b>1 – Partially Meets:</b> 75-89% <b>0 – Does Not Meet:</b> 0-74%
<b>B.3.2</b>	Outcome	All reading questions are text-dependent and focus on the central ideas and important particulars of the text, rather than on superficial or peripheral concepts.	Metrics Auto-Calculated: <ul style="list-style-type: none"> <li>▪ Total reading items</li> <li>▪ Total reading score points</li> <li>▪ Number and percent of items aligned to the specifics of the standard</li> <li>▪ Number and percent of the items requiring close reading.</li> <li>▪ Number and percent of the items focusing on central ideas</li> <li>▪ Number and percent of the items requiring direct textual evidence</li> </ul>	Determine the percentage of items that focus on central ideas and important particulars rather than superficial or peripheral concepts. Assign a rating and provide notes under Comments (for each form):  <b>2 – Meets:</b> Nearly all the items focus on central ideas and important particulars <b>1 – Partially Meets:</b> The large majority of items (i.e., three-quarters or more) focus on central ideas and important particulars. <b>0 – Does Not Meet:</b> Less than a large majority of the items focus on central ideas and important particulars.  Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.	<b>2 – Meets:</b> 90-100% <b>1 – Partially Meets:</b> 75-89% <b>0 – Does Not Meet:</b> 0-74%

<b>B.3 Requiring students to read closely and use evidence from texts:</b> Reading assessments consist of test questions or tasks, as appropriate, that demand that students read carefully and deeply and use specific evidence from increasingly complex texts to obtain and defend correct responses.					
	Type	Evidence Descriptors	Location of Evidence	Scoring Guidance	Tentative Cut-Scores
B.3.3	Outcome	All reading questions are text-dependent and assess the depth and specific requirements delineated in the standards at each grade level (i.e., the concepts, topics, and texts specifically named in the grade-level standards).	<ul style="list-style-type: none"> <li>Number and percent of the reading score points requiring direct textual evidence</li> </ul>	<p>Determine the percentage of items that align to the specifics (i.e., the concepts, topics, and texts) of the standards. Assign a rating and provide notes under Comments (for each form):</p> <p><b>2 – Meets:</b> Nearly all items are aligned to the specifics of the standards.</p> <p><b>1 – Partially Meets:</b> The large majority of items (i.e., three-quarters or more) are aligned to the specifics of the standards.</p> <p><b>0 – Does Not Meet:</b> Less than the large majority of the items are aligned to the specifics of the standards.</p> <p>Note: Items must be aligned to a standard; those that are aligned only to cluster headings (e.g., “Key Ideas and Details”, “Craft and Structure”) or Anchor Standards should be assigned a “0” and rated as Does Not Meet to this metric.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p><b>2 – Meets:</b> 90-100%</p> <p><b>1 – Partially Meets:</b> 75-89%</p> <p><b>0 – Does Not Meet:</b> 0-74%</p>
B.3.4	Outcome	<p>Many reading questions require students to directly provide textual evidence in support of their responses. Goals include:</p> <ul style="list-style-type: none"> <li>A majority of reading score points is devoted to questions that ask students to directly provide textual evidence in support of their responses (e.g., constructed-response and/or two-part evidence-based selected-response item formats).</li> </ul>		<p>Determine the percentage of reading score points that are based on items requiring direct, rather than indirect, use of textual evidence. Assign a rating and provide notes under Comments (for each form):</p> <p><b>2 – Meets:</b> More than half of the reading score points are based on items requiring direct use of textual evidence.</p> <p><b>1 – Partially Meets:</b> Nearly half of the score points are based on items requiring direct use of textual evidence.</p> <p><b>0 – Does Not Meet:</b> Less than one-third of the score points are based on items requiring direct use of textual evidence.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p><b>2 – Meets:</b> 51-100%</p> <p><b>1 – Partially Meets:</b> 33-50%</p> <p><b>0 – Does Not Meet:</b> 0-32%</p>

<b>B.3 Requiring students to read closely and use evidence from texts:</b> Reading assessments consist of test questions or tasks, as appropriate, that demand that students read carefully and deeply and use specific evidence from increasingly complex texts to obtain and defend correct responses.					
	<b>Type</b>	<b>Evidence Descriptors</b>	<b>Location of Evidence</b>	<b>Scoring Guidance</b>	<b>Tentative Cut-Scores</b>
<b>B.3.5</b>	Generalizability	<p>Item specifications require all reading questions to be text-dependent. They require that reading questions:</p> <ul style="list-style-type: none"> <li>• Arise from and require close reading and analysis of text;</li> <li>• Focus on the central ideas and important particulars of the text, rather than on superficial or peripheral concepts; and</li> <li>• Assess the depth and specific requirements delineated in the standards at each grade level – i.e., the concepts, topics, and texts specifically named in the grade-level standards.</li> </ul>	Evidence: Test blueprints and/or other documents identified by the program.	<p>Rate the extent to which the documentation matches the expected percentage of reading items that require close reading, focusing on central ideas, and aligned to the specifics of the standards. Assign a score and provide notes under Comments:</p> <p><b>2 – Meets:</b> Documentation outlines expectations for items to require close reading AND to focus on central ideas and important particulars, AND align to the specifics of the standards.</p> <p><b>1 – Partially Meets:</b> Documentation outlines expectations for only two of the three emphases mentioned above.</p> <p><b>0 – Does Not Meet:</b> Documentation outlines expectations for one or none of the emphases mentioned above.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p><b>2 – Meets:</b> All three</p> <p><b>1 – Partially Meets:</b> Two of three</p> <p><b>0 – Does Not Meet:</b> One of three</p>
<b>B.3.6</b>	Generalizability	<p>Test blueprints or other program documents require that a majority of reading score points be devoted to questions that ask students to directly provide textual evidence in support of their responses (e.g., constructed-response and/or two-part evidence-based selected-response item formats).</p>	Evidence: Test blueprints and/or other documents identified by the program.	<p>Rate the extent to which the documentation matches the expected percentage of reading score points that are based on items requiring direct, rather than indirect, use of textual evidence. Assign a score and provide notes under Comments:</p> <p><b>2 – Meets:</b> Documentation indicates that more than half of the reading score points should be based on items requiring direct use of textual evidence.</p> <p><b>1 – Partially Meets:</b> Documentation indicates that half or less of score points should be based on items requiring direct use of textual evidence.</p> <p><b>0 – Does Not Meet:</b> Documentation indicates that less than one-third of the score points should be based on items requiring direct use of textual evidence.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p><b>2 – Meets:</b> 51-100%</p> <p><b>1 – Partially Meets:</b> 33-50%</p> <p><b>0 – Does Not Meet:</b> 0-32%</p>

<b>B.4 Requiring a range of cognitive demand:</b> The assessments require all students to demonstrate a range of higher-order, analytical thinking skills in reading and writing based on the depth and complexity of college- and career-ready standards, allowing robust information to be gathered for students with varied levels of achievement.					
	Type	Evidence Descriptors	Location of Evidence	Scoring Guidance	Tentative Cut-Offs
B.4.1	Outcome	<p>The distribution of cognitive demand for each grade level and content area is sufficient to assess the depth and complexity of the standards, as evidenced by use of a generic taxonomy (e.g., Webb’s Depth of Knowledge [DoK]) or, preferably, classifications specific to the discipline and drawn from the requirements of the standards themselves and item response modes, such as the:</p> <ul style="list-style-type: none"> <li>• Complexity of the text on which an item is based;</li> <li>• Range of textual evidence an item requires (how many parts of text[s] students must locate and use to response to the item correctly);</li> <li>• Level of inference required; and</li> <li>• Mode of student response (e.g., selected-response, constructed-response).</li> </ul>	<p>Evidence: Test forms Specific metadata from assessment program:</p> <ul style="list-style-type: none"> <li>▪ Point value of item</li> <li>▪ Assigned CCSS alignment (multiple standards shown, if applicable)</li> <li>▪ If program uses Webb, assigned item DoK</li> <li>▪ If program does not use Webb, assigned item cognitive demand level</li> </ul> <p>Coding Sheets:</p> <ul style="list-style-type: none"> <li>▪ By Standard: primary DoK, secondary DoK, tertiary DoK, quaternary DoK.</li> <li>▪ By item: Indicate DoK</li> </ul> <p>Metrics Auto-Calculated: For each test form:</p> <ul style="list-style-type: none"> <li>▪ Number and percent of standards at each of the DoK levels</li> <li>▪ DoK Index = comparing the percentage of score points for items at each DoK level with the percentage of standards at that DoK level, identifying whichever is less, and summing the percentages of the minima</li> <li>▪ DoK Index averaged across both test forms.</li> </ul>	<p>Determine the extent to which the distribution of cognitive demand reflects the cognitive demand of the standards. Assign a score, and provide notes under Comments (for each form).</p> <p><b>2 –Meets:</b> The distribution of cognitive demand of the assessment matches the distribution of cognitive demand of the standards as a whole, AND matches the higher cognitive demand (DoK 3+) of the standards.</p> <p><b>1 – Partially Meets:</b> The distribution of cognitive demand of the assessment partially matches the distribution of cognitive demand of the standards as a whole AND matches the moderate cognitive demand (DoK 2+) of the standards.</p> <p><b>0 – Does Not Meet:</b> The distribution of cognitive demand of the assessment does not match the distribution of cognitive demand of the standards OR has a much higher proportion of low cognitive demand than found in the standards. Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p><b>2 – Meets:</b></p> <ul style="list-style-type: none"> <li>• The DoK Index is at least 80% AND</li> <li>• the percentage of score points associated with DoK3+ items is no more than 10% less than the percentage of standards that are DoK3+.</li> </ul> <p><b>1 – Partially Meets:</b></p> <ul style="list-style-type: none"> <li>• The DoK Index is at least 60% AND</li> <li>• the percent of DoK1 score points is no more than 20% higher than the percentage of standards that are DoK1.</li> </ul> <p><b>0 – Does Not Meet:</b></p> <ul style="list-style-type: none"> <li>• The DoK Index is less than 60% OR</li> <li>• the percent of DoK1 score points is more than 20% greater than the percentage of standards that are DoK1.</li> </ul>

<b>B.4 Requiring a range of cognitive demand:</b> The assessments require all students to demonstrate a range of higher-order, analytical thinking skills in reading and writing based on the depth and complexity of college- and career-ready standards, allowing robust information to be gathered for students with varied levels of achievement.					
	Type	Evidence Descriptors	Location of Evidence	Scoring Guidance	Tentative Cut-Offs
B.4.2	Generalizability	<p>Assessment program has established a definition and procedures for evaluating cognitive demand for assessment items for each grade level and content area that reflects research literature and best practices such as a generic taxonomy (e.g., Webb’s Depth of Knowledge [DoK]) or preferably, classifications specific to the discipline and drawn from the requirements of the standards themselves and item response modes, such as the:</p> <ul style="list-style-type: none"> <li>• Complexity of the text on which an item is based;</li> <li>• Range of textual evidence an item requires (how many parts of text[s] students must locate and use to response to the item correctly);</li> <li>• Level of inference required; and</li> <li>• Mode of student response (e.g., selected-response, constructed-response).</li> </ul>	Evidence: Test blueprints and/or other documents identified by the program.	<p>Rate the extent to which the documentation specifies that the distribution of cognitive demand reflects the cognitive demand of the standards. Assign a score and record notes under Comments.</p> <p><b>2 –Meets:</b> Documentation indicates a research-based definition of cognitive demand, a way of operationalizing cognitive demand at the item level, and a rationale for and specification of distribution of cognitive demand for each test form. The distribution of cognitive demand specified matches the distribution of cognitive demand of the standards as a whole. AND matches the higher cognitive demand of the standards.</p> <p><b>1 – Partially Meets:</b> Documentation indicates a definition of cognitive demand, a way of operationalizing cognitive demand at the item level, and a rationale for and specification of distribution of cognitive demand for each test form. The distribution of cognitive demand specified partially matches the distribution of cognitive demand of the standards as a whole AND matches a moderate cognitive demand of the standards.</p> <p><b>0 – Does Not Meet:</b> Documentation does not indicate a definition of cognitive demand, a way of operationalizing cognitive demand at the item level, or a rationale for and specification of distribution of cognitive demand for each test form. The distribution of cognitive demand specified does not match the distribution of cognitive demand of the standards OR does not match the higher or moderate cognitive demands of the standards.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p><b>2 – Meets:</b></p> <ul style="list-style-type: none"> <li>• If the program uses Webb, the DoK Index is at least 80% AND</li> <li>• the percentage of score points associated with DoK3+ items is no more than 10% less than the percentage of standards that are DoK3+.</li> <li>• If the program uses a measure other than Webb, the definitions, rationales, etc. are appropriate for an assessment program (e.g., specific enough to guide item development and test construction) and the specified distribution of cognitive demand of items on a test form matches the standards as a whole and for the higher demand items/standards.</li> </ul> <p><b>1 – Partially Meets:</b></p> <ul style="list-style-type: none"> <li>• If the program uses Webb, the DoK Index is at least 60% AND</li> <li>• the percent of DoK1 score points is no more than 20% higher than the percentage of standards that are DoK1.</li> <li>• If the program uses a measure other than Webb, the definitions, rationales, etc. are appropriate and the specified distributions of cognitive demand of items on</li> </ul>



**B.4 Requiring a range of cognitive demand:** The assessments require all students to demonstrate a range of higher-order, analytical thinking skills in reading and writing based on the depth and complexity of college- and career-ready standards, allowing robust information to be gathered for students with varied levels of achievement.

	Type	Evidence Descriptors	Location of Evidence	Scoring Guidance	Tentative Cut-Offs
					<p>a test form partially matches the standards as a whole and the lower demand items are not significantly disproportional.</p> <p><b>0 – Does Not Meet:</b></p> <ul style="list-style-type: none"> <li>• If the program uses Webb, the DoK Index is less than 60% OR</li> <li>• the percent of DoK1 score points is more than 20% greater than the percentage of standards that are DoK1.</li> <li>• If the program uses a measure other than Webb, the definitions, rationales, etc. are not appropriate for an assessment program (e.g., too vague to guide item development or test construction) or the specified distribution of cognitive demand of items on a test form does not match that of the standards as a whole or the lower demand items are significantly more than what is in the standards.</li> </ul>

<b>B.5 Assessing writing:</b> Assessments emphasize writing tasks that require students to engage in close reading and analysis of texts so that students can demonstrate college- and career-ready abilities.						
Type	Evidence Descriptors	Location of Evidence	Scoring Evidence	Tentative Cut-Offs		
B.5.1	Outcome	<p>Writing tasks reflect the types of writing that will prepare students for the work required in college and the workplace, balancing expository, persuasive/argument, and narrative writing. At higher grade levels, the balance shifts toward more exposition and argument.</p> <p>Goals include:</p> <ul style="list-style-type: none"> <li>taking all forms of the test together, writing tasks are approximately one-third each exposition, argument, and narrative (some tasks may represent blended structures), with the balance shifting toward more exposition and argument at the higher grade levels.</li> </ul>	<p>Evidence: Test forms, meta-data.</p> <p>Specific metadata from assessment program:</p> <ul style="list-style-type: none"> <li>Assigned CCSS alignment (and secondary alignment(s), if any)</li> <li>Point value of item</li> <li>Chart indicating types of writing assessed at each grade level in the grade band</li> </ul> <p>Coding Sheet:</p> <ul style="list-style-type: none"> <li>What type of writing is called for? (Expository; Persuasive/argumentative; Narrative; Blended)</li> </ul> <p>Coding Sheet Auto calculation:</p> <ul style="list-style-type: none"> <li>Total number of writing items</li> <li>Number and percent of CRs requiring expository writing</li> <li>Number and percent of CRs requiring persuasive/argumentative writing</li> <li>Number and percent of CRs requiring narrative writing</li> <li>Number and percent of CRs requiring blended writing</li> </ul>	<p>Determine the percentages of prompts requiring writing to sources. Assign a score and provide notes under Comments:</p> <p><b>For grades 3 -8 and for high school programs that test narrative writing:</b></p> <p><b>2 – Meets:</b> All three writing types are approximately equally represented across all forms in the grade band, allowing blended types to contribute to the distribution</p> <p><b>1 – Partially Meets:</b> Two of the three writing types are represented across all forms in the grade band, allowing blended types to contribute to the distribution.</p> <p><b>0 – Does Not Meet:</b> One of the three writing types is represented across all forms in the grade band.</p> <p>NOTE: If the high school assessments do not include narrative writing, the assessment can still be rated as Meets.</p> <p><b>For high school programs that do NOT include narrative writing:</b></p> <p><b>2 – Meets:</b> Expository and argument writing types are approximately equally represented across all forms in the grade band, allowing blended types to contribute to the distribution</p> <p><b>1 – Partially Meets:</b> Both writing types are represented but one much more heavily than the other</p> <p><b>0 – Does Not Meet:</b> Only one or no writing type (expository OR argument) is represented.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p><b>For grades 3 -8 and for high school programs that test narrative writing:</b></p> <p><b>2 – Meets:</b> 28-38% of each representing exposition, argument, and narrative</p> <p><b>1 – Partially Meets:</b> Two of the three writing types are present and one type is 0%-27%</p> <p><b>0 – Does Not Meet:</b> One type is 100%</p> <p><b>For high school programs that do NOT include narrative writing:</b></p> <p><b>2 – Meets:</b> 40-60% each for expository and argument types.</p> <p><b>1 – Partially Meets:</b> Both expository and argument types are represented, but one writing type accounts for more than 60% of the balance of these two types.</p> <p><b>0 – Does Not Meet:</b> Either expository or argument is not represented, or neither is represented.</p>	

<b>B.5 Assessing writing:</b> Assessments emphasize writing tasks that require students to engage in close reading and analysis of texts so that students can demonstrate college- and career-ready abilities.					
	Type	Evidence Descriptors	Location of Evidence	Scoring Evidence	Tentative Cut-Offs
B.5.2	Outcome	Tasks (including narrative tasks) require students to confront text or other stimuli directly, to draw on textual evidence, and to support valid inferences from text or stimuli.	<p>Evidence: Test forms, meta-data</p> <p>Specific metadata from assessment program:</p> <ul style="list-style-type: none"> <li>Assigned CCSS alignment (and secondary alignment(s), if any)</li> <li>Point value of item.</li> </ul> <p>Coding Sheet:</p> <ul style="list-style-type: none"> <li>Is the writing task text-based? (Y/N)</li> </ul> <p>Coding Sheet Auto calculation:</p> <ul style="list-style-type: none"> <li>Total number of writing items</li> <li>Number and percent of text-based writing tasks</li> </ul>	<p>Determine the percentages of prompts requiring writing to sources. Assign a score and provide notes under Comments (for each form):</p> <p><b>2 – Meets:</b> All writing prompts require writing to sources (are text-based).</p> <p><b>1 – Partially Meets:</b> The large majority (i.e., three-quarters or more) of writing prompts require writing to sources (are text-based).</p> <p><b>0 – Does Not Meet:</b> Fewer than the large majority of writing prompts require writing to sources (are text-based) OR the program does not include writing prompts.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p><b>2 – Meets:</b> 90-100%</p> <p><b>1 – Partially Meets:</b> 75-89%</p> <p><b>0 – Does Not Meet:</b> 0-74%</p>
B.5.3	Generalizability	<p>Test blueprints and/or other specifications specify the distribution of the various writing tasks/types as standards require, and at higher grade levels the balance shifts toward more exposition and argument. Goals include:</p> <ul style="list-style-type: none"> <li>Taking all forms of the test together, writing tasks are approximately one-third each exposition, argument, and narrative (some tasks may represent blended structures), with the balance shifting toward more exposition and argument at the higher grade levels.</li> </ul>	Evidence: Test blueprints and/or other documents identified by the program.	<p>Determine the degree of match between the specifications of the distribution of the various writing tasks/types and what was expected. Assign a score and provide notes under Comments</p> <p><b>For grades 3 -8 and for high school programs that test narrative writing:</b></p> <p><b>2 –Meets:</b> Documentation indicates that all three writing types are approximately equally represented in the grade band, allowing blended types to contribute to the distribution.</p> <p><b>1 – Partially Meets:</b> Documentation indicates that two of the three writing types are represented in the grade band, allowing blended types to contribute to the distribution</p> <p><b>0 – Does Not Meet:</b> Documentation indicates that one of the three writing types is represented in the grade band.</p>	<p><b>For grades 3 -8 and for high school programs that test narrative writing:</b></p> <p><b>2 – Meets:</b> 28-38% of each representing exposition, argument, and narrative</p> <p><b>1 – Partially Meets:</b> Two of the three writing types are present and one type is 0%-27%</p> <p><b>0 – Does Not Meet:</b> One type is 100%</p> <p><b>For high school programs that do NOT include narrative writing:</b></p>

<b>B.5 Assessing writing:</b> Assessments emphasize writing tasks that require students to engage in close reading and analysis of texts so that students can demonstrate college- and career-ready abilities.					
	Type	Evidence Descriptors	Location of Evidence	Scoring Evidence	Tentative Cut-Offs
				<p>NOTE: If the high school assessments do not include narrative writing, the assessment can still be rated as aligned.</p> <p><b>For high school programs that do NOT include narrative writing:</b></p> <p><b>2 – Meets:</b> Documentation indicates that expository and argument writing types should be approximately equally represented in the grade band, allowing blended types to contribute to the distribution</p> <p><b>1 – Partially Meets:</b> Documentation indicates that both writing types should be represented but one much more heavily than the other (i.e., one writing type accounts for more than 70% of the balance) OR no balance between the two is outlined.</p> <p><b>0 – Does Not Meet:</b> Documentation indicates that only one writing type (expository OR argument) should be represented.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p><b>2 – Meets:</b> 40-60% each for expository and argument types.</p> <p><b>1 – Partially Meets:</b> Both expository and argument types are represented, but one writing type accounts for more than 60% of the balance of these two types.</p> <p><b>0 – Does Not Meet:</b> Either expository or argument is not represented, or neither is represented.</p>
<b>B.5.4</b>	Generalizability	Item and test specifications require students to confront text or other stimuli directly, to draw on textual evidence, and to support valid inferences from text or stimuli.	Evidence: Test blueprints and/or other documents identified by the program.	<p>Determine the degree of match between the specifications of requiring students to confront text or other stimuli directly, to draw on textual evidence, and to support valid inference from text what was expected. Assign a score and provide notes under Comments.</p> <p><b>2 –Meets:</b> Documentation indicates that all writing prompts require writing to sources (are text-based).</p> <p><b>1 – Partially Meets:</b> Documentation indicates that the large majority (i.e., three-quarters or more) of writing prompts require writing to sources (are text-based).</p>	<p><b>2 – Meets:</b> 90-100%</p> <p><b>1 – Partially Meets:</b> 75-89%</p> <p><b>0 – Does Not Meet:</b> 0-74%</p>

**B.5 Assessing writing:** Assessments emphasize writing tasks that require students to engage in close reading and analysis of texts so that students can demonstrate college- and career-ready abilities.

	Type	Evidence Descriptors	Location of Evidence	Scoring Evidence	Tentative Cut-Offs
				<p><b>0 – Does Not Meet:</b> Documentation indicates that fewer than the large majority of writing prompts require writing to sources (are text-based) OR the program does not include writing prompts.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	

<b>B.6 Emphasizing vocabulary and language skills:</b> The assessments require students to demonstrate proficiency in the use of language, including vocabulary and conventions.					
	Type	Evidence Descriptors	Location of Evidence	Scoring Guidelines	Tentative Cut-Offs
B.6.1	Outcome	Vocabulary items reflect requirements for college and career readiness, including focusing on general academic (tier 2) words; asking students to use context to determine meaning; and assessing words that are important to the central ideas of the text.	<p>Evidence: Test forms, meta-data</p> <p>Specific metadata from assessment program:</p> <ul style="list-style-type: none"> <li>▪ Point value for item</li> <li>▪ Primary CCSS alignment</li> <li>▪ Any Secondary CCSS alignment</li> </ul> <p>Coding Sheet:</p> <ul style="list-style-type: none"> <li>• Does the item test a Tier 2 Academic word or phrase? (Y/N)</li> <li>• Does the item test a word central to the understanding of the text? (Y/N)</li> <li>• Does the tested word require use of context? (Y/N)</li> </ul> <p>Coding Sheet Auto calculation:</p> <ul style="list-style-type: none"> <li>• Total vocabulary items</li> <li>• Total vocabulary points</li> <li>• Number and percent of items testing Tier 2 words or phrases</li> <li>• Number and percent of vocabulary items testing words/phrases central to the text</li> <li>• Number and percent of vocabulary items requiring context</li> <li>• Number and percent of vocabulary items testing Tier 2 words or phrases AND requiring context</li> </ul>	<p>Determine the percentage of vocabulary items that focus on tier 2 words, require use of context, and assess words important to central ideas. Assign a score and provide notes under Comments (for each form):</p> <p><b>2 – Meets:</b> The large majority of vocabulary items (i.e., three quarters or more) focuses on tier 2 words AND requires use of context and more than half assess words important to central ideas.</p> <p><b>1 – Partially Meets:</b> At least half of vocabulary items focus on tier 2 words AND require use of context and/or nearly half assess words important to central ideas or in other ways does not quality for 2 or 0.</p> <p><b>0 – Does Not Meet:</b> Less than half of vocabulary items focus on tier 2 words AND require use of context or less than one-third assess words important to central ideas.</p> <p>Note: If less than one-third of vocabulary items assess words that are important to central ideas in the passage, the rating should be 0, regardless of other item characteristics.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p><b>2 – Meets:</b> 75-100% Tier 2; 51% -100% Central</p> <p><b>1 – Partially Meets:</b> 50-75% Tier 2; 33-50% Central</p> <p><b>0 – Does Not Meet:</b> 0-49% Tier 2; 0-32% Central</p>

<b>B.6 Emphasizing vocabulary and language skills:</b> The assessments require students to demonstrate proficiency in the use of language, including vocabulary and conventions.					
	Type	Evidence Descriptors	Location of Evidence	Scoring Guidelines	Tentative Cut-Offs
B.6.2	Outcome	<p>Language is assessed within writing assessments as part of the scoring rubric, or it is assessed with test items that specifically address language skills.</p> <p>Language assessments reflect requirements for college and career readiness by mirroring real-world activities (e.g., actual editing or revision, actual writing); and focusing on common student errors and those conventions most important for readiness.</p>	<p>Evidence: Test forms, meta-data, and writing rubric</p> <p>Specific metadata from assessment program:</p> <ul style="list-style-type: none"> <li>▪ Assigned CCSS alignment (and secondary alignment(s), if any)</li> <li>▪ Score points for each item</li> </ul> <p>Coding Sheet:</p> <ul style="list-style-type: none"> <li>• Does item mirror real-world activities? (Y/N)</li> <li>• Does item test conventions most important for readiness (see CCSS Language Skills Progression Chart)? (Y/N)</li> <li>• Does the item focus on common student errors? (Y/N)</li> </ul> <p>Coding Sheet Auto calculation:</p> <ul style="list-style-type: none"> <li>• Total language items</li> <li>• Total language score points</li> <li>• Number and percent of reading items that mirror real-world activities</li> <li>• Number and percent of items that test conventions most important for readiness</li> <li>• Number and percent of items that focus on common student errors</li> </ul>	<p>Determine the percentage of items in the language skills component that mirror real-world activities, focus on common errors, and emphasize the conventions most important for readiness. Assign a rating and provide notes under Comments (for each form):</p> <p><b>2 – Meets:</b> A large majority (i.e., three-quarters or more) of the items in the language skills component and/or scored with a writing rubric mirror real-world activities, focus on common errors, and emphasize the conventions most important for readiness</p> <p><b>1 – Partially Meets:</b> At least half of the items in the language component and/or scored with a writing rubric mirror real-world activities, focus on common errors, and emphasize the conventions most important for readiness.</p> <p><b>0 – Does Not Meet:</b> Less than half of the items in the language skills component mirror real-world activities, focus on common errors, and emphasize the conventions most important for readiness.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p><b>2 – Meets:</b> 75-100%</p> <p><b>1 – Partially Meets:</b> 50-74%</p> <p><b>0 – Does Not Meet:</b> 0-49%</p>

<b>B.6 Emphasizing vocabulary and language skills:</b> The assessments require students to demonstrate proficiency in the use of language, including vocabulary and conventions.					
	Type	Evidence Descriptors	Location of Evidence	Scoring Guidelines	Tentative Cut-Offs
<b>B.6.3</b>	Outcome	Assessments place sufficient emphasis on vocabulary (i.e., a significant percentage of the score points is devoted to these skills)	<p>Evidence: Test forms, metadata</p> <p>Specific metadata from assessment program:</p> <ul style="list-style-type: none"> <li>Assigned CCSS alignment (and secondary alignment(s), if any)</li> <li>Score points for each item</li> </ul> <p>Coding Sheet Auto calculation:</p> <ul style="list-style-type: none"> <li>Number and percent of score points devoted to assessing vocabulary</li> </ul>	<p>Determine the percentage of score points devoted to assessing vocabulary to support sufficient emphasis. Assign a score and provide notes under Comments (for each form):</p> <p><b>2 – Meets:</b> Vocabulary is reported as a subscore OR at least 13% of score points are devoted to assessing vocabulary</p> <p><b>1 – Partially Meets:</b> At least 10% of score points are devoted to assessing vocabulary</p> <p><b>0 – Does Not Meet:</b> Less than 10% of points are devoted to assessing vocabulary</p> <p>.Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p><b>2 – Meets:</b> Vocabulary is reported as a subscore OR <math>\geq</math> 13% of score points</p> <p><b>1 – Partially Meets:</b> 10 -12% of score points</p> <p><b>0 – Does Not Meet:</b> 0 to 9% of score points</p>
<b>B.6.4</b>	Outcome	Assessments place sufficient emphasis on language skills (i.e., a significant percentage of the score points is devoted to these skills)	<p>Evidence: Test forms, metadata, and writing rubric</p> <p>Specific metadata from assessment program:</p> <ul style="list-style-type: none"> <li>Assigned CCSS alignment (and secondary alignment(s), if any)</li> <li>Score points for each item</li> </ul> <p>Coding Sheet Auto calculation:</p> <ul style="list-style-type: none"> <li>Number and percentage of score points devoted to assessing language.</li> </ul>	<p>If the program includes a language skills component, use the Item Coding Sheet to determine the number and percentage of score points devoted to assessing language. For all programs, use the rubric for the writing test to determine the percentage of score points devoted to assessing language in order to support sufficient emphasis. Assign a score and provide notes under Comments (for each form):</p> <p><b>2 – Meets: Language skills are reported as a subscore OR</b> at least 13% of score points are devoted to assessing language skills (language skills items + score points devoted to assessing language in the writing rubric).</p> <p><b>1 – Partially Meets:</b> At least 10% of score points are devoted to assessing language skills</p> <p><b>0 – Does Not Meet:</b> Less than 10% of points are devoted to assessing language skills</p>	<p><b>2 – Meets: Language skills are reported as a subscore OR</b> <math>\geq</math>13% of score points</p> <p><b>1 – Partially Meets:</b> 10-12% of score points</p> <p><b>0 – Does Not Meet:</b> 0 to 9% of score points</p>



<b>B.6 Emphasizing vocabulary and language skills:</b> The assessments require students to demonstrate proficiency in the use of language, including vocabulary and conventions.					
	Type	Evidence Descriptors	Location of Evidence	Scoring Guidelines	Tentative Cut-Offs
				Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.	
<b>B.6.5</b>	Generalizability	<p>Item specifications require that vocabulary items reflect requirements for college and career readiness, including:</p> <ul style="list-style-type: none"> <li>• Focusing on general academic (tier 2) words;</li> <li>• Asking students to use context to determine meaning; and</li> <li>• Assessing words that are important to the central ideas of the text.</li> </ul>	Evidence: Test blueprints and/or other documents identified by the program.	<p>Determine the percentage of vocabulary items representing tier 2 words and words important to central ideas in the specifications of vocabulary items. Assign a score and provide notes under Comments:</p> <p><b>2 – Meets:</b> Documentation indicates that the large majority (i.e., three-quarters or more) of vocabulary items should focus on tier 2 words AND require use of context and more than half should assess words important to central ideas.</p> <p><b>1 – Partially Meets:</b> Documentation indicates that at least half of vocabulary items should focus on tier 2 words AND should require use of context and/or nearly half should assess words important to central ideas.</p> <p><b>0 – Does Not Meet:</b> Documentation indicates that less than half of vocabulary items should focus on tier 2 words AND should require use of context; OR less than one-third should assess words important to central ideas.</p> <p>Note: If less than one-third of vocabulary items assess words that are important to central ideas in the passage, the rating should be 0, regardless of other item characteristics.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p><b>2 – Meets:</b> 75-100% tier 2 and require use of context; and 51% -100% Central</p> <p><b>1 – Partially Meets:</b> 50-74% tier 2 and require use of context; and/or 33-50% Central</p> <p><b>0 – Does Not Meet:</b> 0-49% tier 2 and require use of context; 0-32% Central</p>

<b>B.6 Emphasizing vocabulary and language skills:</b> The assessments require students to demonstrate proficiency in the use of language, including vocabulary and conventions.					
	Type	Evidence Descriptors	Location of Evidence	Scoring Guidelines	Tentative Cut-Offs
<b>B.6.6</b>	Generalizability	<p>Item specifications require that language is assessed within writing assessments as part of the scoring rubric, or it is assessed with test items that specifically address language skills. Language assessments reflect requirements for college and career readiness by:</p> <ul style="list-style-type: none"> <li>• Mirroring real-world activities (e.g., actual editing or revision, actual writing); and</li> <li>• Focusing on common student errors and those conventions most important for readiness.</li> </ul>	Evidence: Test blueprints and/or other documents identified by the program.	<p>Determine the percent of items mirroring real-world activities, focusing on common errors, and emphasizing the conventions most important for readiness in the specifications. Assign a score and provide notes under Comments:</p> <p><b>2 – Meets:</b> Documentation indicates that the large majority (i.e., three-quarters or more) of the items in the language skills component and/or scored with a writing rubric should mirror real-world activities, focus on common errors, and emphasize the conventions most important for readiness.</p> <p><b>1 – Partially Meets:</b> Documentation indicates that at least half of the items in the language component and/or scored with a writing rubric should mirror real-world activities, focus on common errors, and emphasize the conventions most important for readiness.</p> <p><b>0 – Does Not Meet:</b> Documentation indicates that less than half of the items in the language skills component should mirror real-world activities, focus on common errors, and emphasize the conventions most important for readiness.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p><b>2 – Meets:</b> 75-100%</p> <p><b>1 – Partially Meets:</b> 50-74%</p> <p><b>0 – Does Not Meet:</b> 0-49%</p>
<b>B.6.7</b>	Generalizability	Test blueprints and other specifications for each grade level place sufficient emphasis on vocabulary (i.e., a significant percentage of the score points is devoted to these skills)	Evidence: Test blueprints and/or other documents identified by the program.	<p>Determine the percentage of score points associated with vocabulary to support sufficient emphasis and provide notes under Comments:</p> <p><b>2 – Meets:</b> Documentation indicates that vocabulary is reported as a subscore OR at least 13% of score points should be devoted to assessing vocabulary.</p> <p><b>1 – Partially Meets:</b> Documentation indicates that at least 10% of score points should be devoted to</p>	<p><b>2 – Meets:</b> Vocabulary is reported as a subscore or <math>\geq</math> 13% of score points</p> <p><b>1 – Partially Meets:</b> 10=12% of score points</p> <p><b>0 – Does Not Meet:</b> 0 to 9% of score points</p>

<b>B.6 Emphasizing vocabulary and language skills:</b> The assessments require students to demonstrate proficiency in the use of language, including vocabulary and conventions.					
	Type	Evidence Descriptors	Location of Evidence	Scoring Guidelines	Tentative Cut-Offs
				assessing vocabulary. <b>0 – Does Not Meet:</b> Documentation indicates that less than 10% or score points should be devoted to assessing vocabulary.  Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.	
<b>B.6.8</b>	Generalizability	<ul style="list-style-type: none"> <li>Assessments place sufficient emphasis on vocabulary and language skills (i.e., a significant percentage of the score points is devoted to these skills)</li> </ul>	Evidence: Test blueprints and/or other documents identified by the program.	Determine the percentage of score points devoted to language skills and provide notes under Comments:  <b>2 – Meets:</b> Documentation indicates that language skills are reported as a subscore OR at least 13% of score points should be devoted to assessing language skills (language skills items + score points devoted to assessing language in the writing rubric). <b>1 – Partially Meets:</b> Documentation indicates that at least 10% of score points should be devoted to assessing language skills. <b>0 – Does Not Meet:</b> Documentation indicates that less than 10% of or fewer points should be devoted to assessing language skills.  Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.	<b>2 – Meets:</b> Language skills are reported as a subscore OR $\geq$ 13% of score points <b>1 – Partially Meets:</b> 10-12% of score points <b>0 – Does Not Meet:</b> Less than 10% of score points

**B.7 Assessing research and inquiry:** The assessments require students to demonstrate research and inquiry skills, demonstrated by the ability to find, process, synthesize, organize, and use information from sources.

	Type	Evidence Descriptors	Location of Evidence	Scoring Guidelines	Tentative Cut-Offs
<b>B.7.1</b>	Outcome	<p>Test items assessing research and inquiry mirror real world activities and require students to analyze, synthesize, organize, and use information from sources. Goals include:</p> <ul style="list-style-type: none"> <li>• Research tasks require writing to sources, including analyzing, selecting, and organizing evidence from more than one source, and often from sources in diverse formats</li> </ul>	<p>Evidence: Test forms, meta-data</p> <p>Specific metadata from assessment program:</p> <ul style="list-style-type: none"> <li>▪ Point value</li> <li>▪ Grade level</li> <li>▪ Primary assigned CCSS alignment</li> <li>▪ Any secondary CCSS alignment</li> </ul> <p>Coding Sheet:</p> <ul style="list-style-type: none"> <li>• Does item require analysis, synthesis, and/or organization of information (mirroring real-world activities)? (Y/N)</li> </ul> <p>Coding Sheet Auto calculation:</p> <ul style="list-style-type: none"> <li>• Total research items</li> <li>• Total research score points</li> <li>• Number and percent of items mirroring real world activities</li> <li>• Number and percent of items devoted to research</li> <li>• Number and percent of points devoted to research</li> </ul>	<p>Determine the percentage of research skills items that require analysis, synthesis, and/or organization of information. Assign a score and provide notes under Comments (for each form):</p> <p><b>2 – Meets:</b> The large majority (i.e., three-quarters or more) of the research items require analysis, synthesis, and/or organization of information.</p> <p><b>1 – Partially Meets:</b> More than half of the research items require analysis, synthesis, and/or organization of information.</p> <p><b>0 – Does Not Meet:</b> Half or less than half of research items require analysis, synthesis, and/or organization of information</p> <p>NOTES: If there is no research component, score this as 0. If the assessment offers paired nonfictional passages with a writing task, count that section of the test as research.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p><b>2 – Meets:</b> 75-100%</p> <p><b>1 – Partially Meets:</b> 51-74%</p> <p><b>0 – Does Not Meet:</b> 0-50%</p>

<b>B.7 Assessing research and inquiry:</b> The assessments require students to demonstrate research and inquiry skills, demonstrated by the ability to find, process, synthesize, organize, and use information from sources.					
	Type	Evidence Descriptors	Location of Evidence	Scoring Guidelines	Tentative Cut-Offs
B.7.2	Generalizability	<p>Test blueprints and other specifications as well as exemplar test items for each grade level are provided, demonstrating the expectations below are met. Goals include:</p> <ul style="list-style-type: none"> <li>When assessment constraints permit, real or simulated research tasks comprise a significant percentage of score points when all forms of the reading and writing test are considered together.</li> </ul>	Evidence: Test blueprints and/or other documents identified by the program.	<p>Determine the percentage of score points assessing real or simulated research tasks. Assign a score and provide notes under Comments:</p> <p><b>2 – Meets:</b> Program reports a research score or otherwise demonstrates that research is significant.</p> <p><b>1 – Partially Meets:</b> Program includes research items should be assessed but these are not reported or program does not indicate research is significant.</p> <p><b>0 – Does Not Meet:</b> No research items are specified to be included.</p> <p>Note: A research item, at a minimum, includes paired nonfiction passages with a writing task.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p><b>2 – Meets:</b> Program reports a research score or otherwise demonstrates that research is significant.</p> <p><b>1 – Partially Meets:</b> Program includes research items should be assessed but these are not reported or indicates research is not significant.</p> <p><b>0 – Does Not Meet:</b> No research items are specified to be included.</p>
B.7.3	Generalizability	<p>Item specifications and/or other ancillary documents specify that test items assessing research and inquiry mirror real world activities and require students to analyze, synthesize, organize, and use information from sources. Goals include:</p> <ul style="list-style-type: none"> <li>Research tasks require writing to sources, including analyzing, selecting, and organizing evidence from more than one source, and often from sources in diverse formats.</li> </ul>	Evidence: Test blueprints and/or other documents identified by the program.	<p>Determine the percentage of test items assessing research and inquiry mirroring real world activities. Assign a score and provide notes under Comments:</p> <p><b>2 – Meets:</b> Documentation indicates that the large majority (i.e., three-quarters or more) of the research items require analysis, synthesis, and/or organization of information.</p> <p><b>1 – Partially Meets:</b> Documentation indicates that more than half of the research items require analysis, synthesis, and/or organization of information</p> <p><b>0 – Does Not Meet:</b> Documentation indicates that half or less than half of research items require analysis, synthesis, and/or organization of information.</p> <p>NOTES: If there is no research component, rate this evidence descriptor as 0.</p>	<p><b>2 – Meets:</b> 75-100%</p> <p><b>1 – Partially Meets:</b> 51-74%</p> <p><b>0 – Does Not Meet:</b> 0-50%</p>

**B.7 Assessing research and inquiry:** The assessments require students to demonstrate research and inquiry skills, demonstrated by the ability to find, process, synthesize, organize, and use information from sources.

	<b>Type</b>	<b>Evidence Descriptors</b>	<b>Location of Evidence</b>	<b>Scoring Guidelines</b>	<b>Tentative Cut-Offs</b>
				<p>If the assessment offers paired nonfictional passages with a writing task, count that section of the test as research.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the "Location of Evidence" column were not available.</p>	

**B.8 Assessing speaking and listening:** Over time, and as assessment advances allow, the assessments measure the speaking and listening communication skills students need for college and career readiness.

	Type	Evidence Descriptors	Location of Evidence	Scoring Guidelines	Tentative Cut-Offs
B.8.1	Outcome	<p>Over time, and as assessment advances allow, the listening skills required for college and career readiness are assessed.</p> <p>Test items assessing listening:</p> <ul style="list-style-type: none"> <li>• Are based on texts and other stimuli that meet the criteria for complexity, range, and quality outlined in criteria B.1 and B.2 above; and</li> <li>• Permit the evaluation of active listening skills (e.g., taking notes on main ideas, elaborating on remarks of others).</li> </ul>	<p>Evidence: Test forms, meta-data</p> <p>Specific metadata from assessment program:</p> <ul style="list-style-type: none"> <li>▪ Assigned CCSS alignment (and any secondary alignment(s), if any)</li> </ul> <p>Coding Sheet:</p> <ul style="list-style-type: none"> <li>• Does listening stimulus meet expectations for quality as outlined in B.1? (Y/N) "B.8.1</li> <li>• Does the listening stimulus meet the expectations for complexity outlined in B2? (Y/N)</li> <li>• Does listening item require active listening? (Y/N)</li> </ul> <p>Coding Sheet Auto calculation:</p> <ul style="list-style-type: none"> <li>• Total listening items</li> <li>• Number and percent of listening items with stimuli that meet B.1 &amp; B.2 expectations for complexity and quality</li> <li>• Number and percent of listening items that require active listening</li> <li>• Number and percent of listening items that require active listening AND with stimuli that meet B.1 &amp; B.2 expectations for complexity and quality</li> </ul>	<p>Determine the percentage of items are based on texts and other stimuli that meet the criteria for complexity, range, and quality outlined in criteria B.1 and B.2 above and require evaluation of active listening skills. Assign a score and provide notes under Comments (for each form).</p> <p><b>2 – Meets:</b> The large majority (i.e., at least three-quarters) of listening items meet the requirements outlined in B.1 and B.2 AND evaluate active listening skills.</p> <p><b>1 – Partially Meets:</b> Many (i.e., at least half) of listening items meet the requirements outlined in B.1 and B.2 AND evaluate active listening skills.</p> <p><b>0 – Does Not Meet:</b> Less than half of the listening items meet the requirements outlined in B.1 and B.2 AND less than half evaluate active listening skills.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p><b>2 – Meets:</b> 75-100%</p> <p><b>1 – Partially Meets:</b> 50-74%</p> <p><b>0 – Does Not Meet:</b> 0-49%</p>

<b>B.8 Assessing speaking and listening:</b> Over time, and as assessment advances allow, the assessments measure the speaking and listening communication skills students need for college and career readiness.					
	Type	Evidence Descriptors	Location of Evidence	Scoring Guidelines	Tentative Cut-Offs
<b>B.8.2</b>	Outcome	<p>Over time, and as assessment advances allow, the speaking skills required for college and career readiness are assessed.</p> <p>Test items assessing speaking:</p> <ul style="list-style-type: none"> <li>Assess students' ability to express well-supported ideas clearly and to probe others' ideas; and</li> <li>Include items that measure students' ability to marshal evidence from research and orally present findings in a performance task.</li> </ul>	<p>Evidence: Test forms, meta-data</p> <p>Specific metadata from assessment program:</p> <ul style="list-style-type: none"> <li>Assigned CCSS alignment</li> </ul> <p>Coding Sheet:</p> <ul style="list-style-type: none"> <li>Does the item assess student's ability to express well supported ideas clearly and to probe other's ideas? (Y/N)</li> <li>Does the item measure students' ability to marshal evidence from research? (Y/N)</li> <li>Does the item measure students' ability to orally present findings? (Y/N)</li> </ul> <p>Coding Sheet Auto calculation:</p> <ul style="list-style-type: none"> <li>Number and percent of speaking items assessing students' ability to express well supported ideas and probe others ideas               <ul style="list-style-type: none"> <li>Number and percent of speaking items that measure students ability to marshal evidence from research.</li> </ul> </li> <li>Number and percent of speaking items that measure students' ability to orally present findings.</li> </ul>	<p>Determine the percentage of items that require students to express well-supported ideas clearly and to probe others' ideas; to marshal evidence from research; and to present findings orally. Assign a score and provide notes under Comments (for each form).</p> <p><b>2 – Meets:</b> The large majority (i.e., at least three-quarters) of speaking items assess students' ability to do all three of these things: express well-supported ideas clearly and to probe others' ideas; AND marshal evidence from research; AND present findings orally in a performance task.</p> <p><b>1 – Partially Meets:</b> Many (at least half) of speaking items assess students' ability to do all three of these things: express well-supported ideas clearly and to probe others' ideas; AND marshal evidence from research; AND present findings orally in a performance task.</p> <p><b>0 – Does Not Meet:</b> Less than half of speaking items assess students' ability to do all three of these things: express well-supported ideas clearly and to probe others' ideas; AND marshal evidence from research; AND present findings orally in a performance task.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the "Location of Evidence" column were not available.</p>	<p><b>2 – Meets:</b> 75-100%</p> <p><b>1 – Partially Meets:</b> 50-74%</p> <p><b>0 – Does Not Meet:</b> 0-49%</p>



**B.8 Assessing speaking and listening:** Over time, and as assessment advances allow, the assessments measure the speaking and listening communication skills students need for college and career readiness.

	Type	Evidence Descriptors	Location of Evidence	Scoring Guidelines	Tentative Cut-Offs
			<ul style="list-style-type: none"> <li>Number and percent of speaking items assessing students ability to express well supported ideas and probe others ideas AND marshal evidence from research and orally present findings.</li> </ul>		
<b>B.8.3</b>	Generalizability	<p>Item specifications and other ancillary documents specify that test items assessing listening reflect current assessment capabilities and constraints.</p> <p>Test items assessing listening:</p> <ul style="list-style-type: none"> <li>Are based on texts and other stimuli that meet the criteria for complexity, range, and quality outlined in criteria B.1 and B.2 above; and</li> <li>Permit the evaluation of active listening skills (e.g., taking notes on main ideas, elaborating on remarks of others).</li> </ul>	Evidence: Test blueprints and/or other specification documents.	<p>Determine the percentage of test items being based on texts and other stimuli that meet the criteria for complexity range, and quality in criteria B.1 and B.2. Assign a score and provide notes under Comments:</p> <p><b>2 – Meets:</b> Documentation indicates the large majority (i.e., at least three-quarters) of listening items should meet the requirements outlined in B.1 and B.2 AND they should evaluate active listening skills.</p> <p><b>1 – Partially Meets:</b> Documentation indicates that at least half of listening items should meet the requirements outlined in B.1 and B.2 AND they should evaluate active listening skills.</p> <p><b>0 – Does Not Meet:</b> Documentation indicates that less than half of the listening items should meet the requirements outlined in B.1 and B.2 AND less than half should evaluate active listening skills.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p><b>2 – Meets:</b> 75-100%</p> <p><b>1 – Partially Meets:</b> 50-74%</p> <p><b>0 – Does Not Meet:</b> 0-49%</p>

**B.8 Assessing speaking and listening:** Over time, and as assessment advances allow, the assessments measure the speaking and listening communication skills students need for college and career readiness.

	Type	Evidence Descriptors	Location of Evidence	Scoring Guidelines	Tentative Cut-Offs
<b>B.8.4</b>	Generalizability	<p>Item specifications and other ancillary documents specify that test items assessing speaking reflect current assessment capabilities and constraints.</p> <p>Test items assessing speaking:</p> <ul style="list-style-type: none"> <li>Assess students' ability to express well-supported ideas clearly and to probe others' ideas; and</li> <li>Include items that measure students' ability to marshal evidence from research and orally present findings in a performance task.</li> </ul>	Evidence: Test blueprints and/or other specification documents.	<p>Determine the percentage of items that require students to express well-supported ideas clearly and to probe others' ideas, marshal evidence from research, and present findings orally. Assign a score and provide notes under Comments:</p> <p><b>2 – Meets:</b> Documentation outlines the expectation that the large majority (i.e., at least three-quarters) of speaking items assess students' ability to do all three of these things: express well-supported ideas clearly and to probe others' ideas; AND marshal evidence from research; AND present findings orally in a performance task.</p> <p><b>1 – Partially Meets:</b> Documentation outlines the expectation that at least half of speaking items assess students' ability to do all three of these things: express well-supported ideas clearly and to probe others' ideas; AND marshal evidence from research; AND orally present findings in a performance task.</p> <p><b>0 – Does Not Meet:</b> Documentation outlines that less than half of speaking items assess students' ability to do all three of these things: express well-supported ideas clearly and to probe others' ideas; AND measure students' ability to marshal evidence from research; AND orally present findings in a performance task.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the "Location of Evidence" column were not available.</p>	<p><b>2 – Meets:</b> 75-100%</p> <p><b>1 – Partially Meets:</b> 50-74%</p> <p><b>0 – Does Not Meet:</b> 0-49%</p>

<b>B.9 Ensuring high-quality items and a variety of item types:</b> High-quality items and a variety of types are strategically used to appropriately assess the standard(s).					
	Type	Evidence Descriptors	Location of Evidence	Scoring Guidance	Tentative Cut-Offs
B.9.1	Outcome	Items are reviewed to ensure that the distribution of item types for each grade level and content area is sufficient to strategically assess the depth and complexity of the standards being addressed. Item types may include, for example, selected-response, two-part evidence-based selected-response, short and extended constructed-response, technology-enhanced, and performance tasks.	<p>Evidence: Test forms, meta-data</p> <p>Specific metadata from assessment program:</p> <ul style="list-style-type: none"> <li>▪ Item type</li> </ul> <p>Coding Sheet:</p> <ul style="list-style-type: none"> <li>▪ Are there 2 or more item types? (Y/N)</li> </ul> <p>Does at least one of the item types require students to generate, rather than select, a response? (Y/N)</p> <p>Coding Sheet auto calculation:</p> <ul style="list-style-type: none"> <li>• Number and percent of multiple choice items</li> <li>• Number and percent of multi-select items</li> <li>• Number and percent of evidence-based selected response items</li> <li>• Number and percent of technology enhanced items (does not require student to generate a response)</li> <li>• Number and percent of constructed/student generated responses</li> <li>• Number and percent of items with other item type</li> <li>• Number and percent of high quality items</li> </ul>	<p>Determine the kinds of item formats used on the operational forms. Assign a score and provide notes under Comments (for each form):</p> <p><b>2 – Meets:</b> At least two item formats are used, including one that requires students to generate, rather than select, a response (i.e., CR, extended writing).</p> <p><b>1 – Partially Meets:</b> At least two formats (but not including CR) are used, including technology-based formats and/or two-part selected response formats.</p> <p><b>0 – Does Not Meet:</b> Only a traditional multiple choice format is used.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p><b>2 – Meets:</b> At least two item formats are used, including one that requires students to generate, rather than select, a response (i.e., CR, extended writing).</p> <p><b>1 – Partially Meets:</b> At least two formats (but not including CR) are used, including technology-based formats and/or two-part selected response formats.</p> <p><b>0 – Does Not Meet:</b> Only a traditional multiple choice format is used.</p>

<b>B.9 Ensuring high-quality items and a variety of item types:</b> High-quality items and a variety of types are strategically used to appropriately assess the standard(s).					
	<b>Type</b>	<b>Evidence Descriptors</b>	<b>Location of Evidence</b>	<b>Scoring Guidance</b>	<b>Tentative Cut-Offs</b>
<b>B.9.2</b>	Outcome	Operational items are reviewed to verify claims of quality, including ensuring the technical quality, alignment to standards, and editorial accuracy of the items.	<p>Evidence: Test forms, meta-data</p> <p>Specific metadata from assessment program:</p> <ul style="list-style-type: none"> <li>▪ Point value of item</li> <li>▪ Assigned CCSS alignment</li> </ul> <p>Coding Sheets:</p> <ul style="list-style-type: none"> <li>▪ Do you agree with the assigned CCSS Alignment? (Y/N)</li> <li>▪ Is there a quality issue with this item? (Y/N)</li> <li>▪ If so, what is the issue? (Select all that apply) <ul style="list-style-type: none"> <li>○ Item may not yield valid evidence of targeted skill</li> <li>○ Item has issues with readability</li> <li>○ Item incorrectly keyed</li> <li>○ Item has unintended correct answer</li> <li>○ Content is inaccurate</li> <li>○ Item has issues with editorial accuracy</li> </ul> </li> </ul> <p>Metrics auto-calculated:</p> <ul style="list-style-type: none"> <li>▪ % of high-quality items</li> <li>▪ % of agreement with given alignment</li> </ul>	<p>Using the provided documentation, determine that there are high-quality items. Assign a score and provide notes under Comments (for each form):</p> <p><b>2 –Meets:</b> All or nearly all operational items reviewed reflect technical quality, alignment to standards, and editorial accuracy.</p> <p><b>1 – Partially Meets:</b> A few operational items reviewed have issues with technical quality, alignment to standards, and/or editorial accuracy.</p> <p><b>0 – Does Not Meet:</b> Enough of the operational items reviewed have issues with technical quality, alignment to standards, and/or editorial accuracy that quality issues significantly impact the ability of the form to measure important constructs.</p> <p>Note: Reviewers may enter comments about the quality of specific items in the Item Worksheet.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p><b>2 – Meets:</b> 95-100% for editorial and technical; 90% for alignment to standards</p> <p><b>1 – Partially Meets:</b> 90-94% for editorial and technical; 80% for alignment to standards</p> <p><b>0 – Does Not Meet:</b> 0-89% for editorial and technical; 0-79% for alignment to standards</p>

<b>B.9 Ensuring high-quality items and a variety of item types:</b> High-quality items and a variety of types are strategically used to appropriately assess the standard(s).					
	<b>Type</b>	<b>Evidence Descriptors</b>	<b>Location of Evidence</b>	<b>Scoring Guidance</b>	<b>Tentative Cut-Offs</b>
<b>B.9.3</b>	Generalizability	Specifications are provided to demonstrate that the distribution of item types for each grade level and content area is sufficient to strategically assess the depth and complexity of the standards being addressed.	Evidence: Test blueprints and/or other documents identified by the program.	<p>Assign a score representing the specification for ensuring high-quality items and a variety of item types; provide notes under Comments:</p> <p><b>2 – Meets:</b> Documentation indicates that at least two item formats should be used, including one that requires students to generate, rather than select, a response (i.e., CR, extended writing).</p> <p><b>1 – Partially Meets:</b> Documentation indicates that at least two formats (but not including CR) should be used, including technology-based formats and/or two-part selected response formats.</p> <p><b>0 – Does Not Meet:</b> Documentation indicates that only a single format should be used, including traditional multiple-choice format.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p><b>2 – Meets:</b> Specifications indicate that at least two item formats should be used, including one that requires students to generate, rather than select, a response (i.e., CR, extended writing).</p> <p><b>1 – Partially Meets:</b> Specifications indicate that at least two formats (but not including CR) should be used, including technology-based formats and/or two-part selected response formats.</p> <p><b>0 – Does Not Meet:</b> Specifications indicate that only a single format should be used, including traditional multiple-choice format.</p>
<b>B.9.4</b>	Generalizability	<p>To support claims of quality, the following are provided in documentation:</p> <ul style="list-style-type: none"> <li>• Rationales for the use of the specific item types;</li> <li>• Specifications showing the proportion of item types on a form;</li> <li>• For constructed response and performance tasks, a scoring plan (e.g., machine-scored, hand-scored, by whom, how trained), scoring rubrics, and sample student work to confirm the validity of the scoring process;</li> </ul> <p>A description of the process used for ensuring the technical quality, alignment to standards, and editorial accuracy of the items.</p>	Evidence: Test blueprints, administration and scoring manuals, QC procedure documents, and/or other documents provided by the program.	<p>Assign a score and provide notes under Comments:</p> <p><b>2 –Meets:</b> Documentation supports claims of the technical quality, alignment to standards, and editorial accuracy.</p> <p><b>1 – Partially Meets:</b> Documentation partially supports claims of the technical quality, alignment to standards, and/or editorial accuracy.</p> <p><b>0 – Does Not Meet:</b> Documentation does not support claims of the technical quality, alignment to standards, and/or editorial accuracy.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p><b>2 –Meets:</b> Documentation supports claims of the technical quality, alignment to standards, and editorial accuracy.</p> <p><b>1 – Partially Meets:</b> Documentation partially supports claims of the technical quality, alignment to standards, and/or editorial accuracy.</p> <p><b>0 – Does Not Meet:</b> Documentation does not support claims of the technical quality, alignment to standards, and/or editorial accuracy.</p>

### SCORING SUMMARY

Criterion	Sub-Criterion	Score		Automatic Criterion-Level Raw Score	Automatic Criterion Score	Group Rating		Automatic Criterion-Level Raw Score	Automatic Criterion Score		
		Form 1	Form 2			Form 1	Form 2				
B.1	Assessing student reading and writing achievement in both ELA and literacy	B.1.1	<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing	Add (0/1/2) ratings from each form and each outcome sub-criterion. Range: 0 to 12	10- 12 = E 7-9 = G 4-6 = L 0-3 = W	<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing	Add (0/1/2) ratings from each form and each outcome sub-criterion. Range: 0 to 12	E G L W	
		B.1.2	<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing			<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing			
		B.1.2	<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing			<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing			
		Comments:									
		B.1.4			(0/1/2) Rating		Indicate degree of confidence: + : Outcome ratings are likely to be seen in other forms = : Neither confident nor pessimistic - : Outcome ratings are unlikely to be seen in other forms ☒ : Documentation missing				
				<input type="checkbox"/> : Missing							
		B.1.5			(0/1/2) Rating						
					<input type="checkbox"/> : Missing						
		B.1.6			(0/1/2) Rating						
			<input type="checkbox"/> : Missing								
Comments:											
B.2	Focusing on complexity of texts	B.2.1	<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing	Add (0/1/2) ratings from each form and each outcome sub-criterion. Range: 0 to 4	4 = E 3 = G 2 = L 0-1 = W	<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing	Add (0/1/2) ratings from each form and each outcome sub-criterion. Range: 0 to 4	E G L W	
		Comments									

Criterion		Sub-Criterion	Score		Automatic Criterion-Level Raw Score	Automatic Criterion Score	Group Rating		Automatic Criterion-Level Raw Score	Automatic Criterion Score
			Form 1	Form 2			Form 1	Form 2		
		<b>B.2.2</b>			(0/1/2) Rating <input type="checkbox"/> : Missing		Indicate degree of confidence: +: Outcome ratings are likely to be seen in other forms =: Neither confident nor pessimistic -: Outcome ratings are unlikely to be seen in other forms ☒: Documentation missing			
		<b>B.2.3</b>			(0/1/2) Rating <input type="checkbox"/> : Missing					
		<b>Comments</b>								
<b>B.3</b>	<b>Requiring students to read closely and use evidence from texts</b>	<b>B.3.1</b>	<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing	Add (0/1/2) ratings from each form and each outcome sub-criterion. Range: 0 to 16	13-16 = E 9-12 = G 5-8 = L 0-4 = W	<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing	Add (0/1/2) ratings from each form and each outcome sub-criterion. Range: 0 to 16	E G L W
		<b>B.3.2</b>	<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing			<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing		
		<b>B.3.3</b>	<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing			<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing		
		<b>B.3.3</b>	<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing			<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing		
		<b>Comments</b>								
		<b>B.3.5</b>			(0/1/2) Rating <input type="checkbox"/> : Missing		Indicate degree of confidence: +: Outcome ratings are likely to be seen in other forms =: Neither confident nor pessimistic -: Outcome ratings are unlikely to be seen in other forms ☒: Documentation missing			
		<b>B.3.6</b>			(0/1/2) Rating <input type="checkbox"/> : Missing					
		<b>Comments</b>								

Criterion		Sub-Criterion	Score		Automatic Criterion-Level Raw Score	Automatic Criterion Score	Group Rating		Automatic Criterion-Level Raw Score	Automatic Criterion Score	
			Form 1	Form 2			Form 1	Form 2			
B.4	Requiring a range of cognitive demand	B.4.1			Add (0/1/2) ratings from each form and each outcome sub-criterion. Range: 0 to 4	4 = E 3 = G 2 = L 0-1 = W			Add (0/1/2) ratings from each form and each outcome sub-criterion. Range: 0 to 4	E G L W	
			<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing			<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing			
		Comments									
		B.4.2			(0/1/2) Rating		Indicate degree of confidence: + : Outcome ratings are likely to be seen in other forms = : Neither confident nor pessimistic - : Outcome ratings are unlikely to be seen in other forms ☒ : Documentation missing				
			<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing						
Comments											
B.5	Assessing writing	B.5.1			Add (0/1/2) ratings from each form and each outcome sub-criterion. Range: 0 to 8	7- 8 = E 5-6 = G 3-4 = L 0-2 = W			Add (0/1/2) ratings from each form and each outcome sub-criterion. Range: 0 to 8	E G L W	
			<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing			<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing			
		B.5.2									
			<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing			<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing			
		Comments									
		B.5.3			(0/1/2) Rating		Indicate degree of confidence: + : Outcome ratings are likely to be seen in other forms = : Neither confident nor pessimistic - : Outcome ratings are unlikely to be seen in other forms ☒ : Documentation missing				
			<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing						
B.5.4			(0/1/2) Rating								
	<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing								
Comments											



Criterion		Sub-Criterion	Score		Automatic Criterion-Level Raw Score	Automatic Criterion Score	Group Rating		Automatic Criterion-Level Raw Score	Automatic Criterion Score																				
			Form 1	Form 2			Form 1	Form 2																						
B.6	Emphasizing vocabulary and language skills	B.6.1			Add (0/1/2) ratings from each form and each outcome sub-criterion. Range: 0 to 16	13-16 = E 9-12 = G 5-8 = L 0-4 = W			Add (0/1/2) ratings from each form and each outcome sub-criterion. Range: 0 to 16	E G L W																				
			<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing			<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing																						
		B.6.2																												
			<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing										<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing															
		B.6.3																												
			<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing															<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing										
		B.6.4																												
			<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing																				<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing					
		Comments																												
		B.6.5																						(0/1/2) Rating					Indicate degree of confidence: + : Outcome ratings are likely to be seen in other forms = : Neither confident nor pessimistic - : Outcome ratings are unlikely to be seen in other forms ☒ : Documentation missing	
			<input type="checkbox"/> : Missing																											
		B.6.6																						(0/1/2) Rating						
<input type="checkbox"/> : Missing																														
B.6.7			(0/1/2) Rating																											
	<input type="checkbox"/> : Missing																													
B.6.8			(0/1/2) Rating																											
	<input type="checkbox"/> : Missing																													
Comments																														

Criterion		Sub-Criterion	Score		Automatic Criterion-Level Raw Score	Automatic Criterion Score	Group Rating		Automatic Criterion-Level Raw Score	Automatic Criterion Score	
			Form 1	Form 2			Form 1	Form 2			
B.7	Assessing research and inquiry	B.7.1	<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing	Add (0/1/2) ratings from each form and each outcome sub-criterion. Range: 0 to 4	4 = E 3 = G 2 = L 0-1 = W	<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing	Add (0/1/2) ratings from each form and each outcome sub-criterion. Range: 0 to 4	E G L W	
		Comments									
		B.7.2			(0/1/2) Rating <input type="checkbox"/> : Missing		Indicate degree of confidence: + : Outcome ratings are likely to be seen in other forms = : Neither confident nor pessimistic - : Outcome ratings are unlikely to be seen in other forms ☒ : Documentation missing				
		B.7.3			(0/1/2) Rating <input type="checkbox"/> : Missing						
		Comments									
B.8	Assessing speaking and listening	B.8.1	<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing	Add (0/1/2) ratings from each form and each outcome sub-criterion. Range: 0 to 8	4 = E 3 = G 2 = L 0-1 = W	<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing	Add (0/1/2) ratings from each form and each outcome sub-criterion. Range: 0 to 4	E G L W	
		Comments									
		B.8.2	<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing			<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing			
		Comments									
		B.8.3			(0/1/2) Rating <input type="checkbox"/> : Missing						
		Comments									
		B.8.4									
						<input type="checkbox"/> : Missing					
		Comments									

Criterion		Sub-Criterion	Score		Automatic Criterion-Level Raw Score	Automatic Criterion Score	Group Rating		Automatic Criterion-Level Raw Score	Automatic Criterion Score	
			Form 1	Form 2			Form 1	Form 2			
<b>B.9</b>	<b>Ensuring high-quality items and a variety of item types</b>	<b>B.9.1</b>			Add (0/1/2) ratings from each form and each outcome sub-criterion. Range: 0 to 8	7-8 = E 5-6 = G 3-4 = L 0-2 = W			Add (0/1/2) ratings from each form and each outcome sub-criterion. Range: 0 to 8	E G L W	
			<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing			<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing			
		<b>B.9.2</b>									
			<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing					
		<b>Comments</b>									
		<b>B.9.2</b>				(0/1/2) Rating		Indicate degree of confidence: +: Outcome ratings are likely to be seen in other forms =: Neither confident nor pessimistic -: Outcome ratings are unlikely to be seen in other forms ☒: Documentation missing			
				<input type="checkbox"/> : Missing							
<b>Comments</b>											

### Cluster Scoring Rules:

The overall rating for the super-criterion should not be higher than the rating for the emphasized criteria. In cases where there is one emphasized criterion (i.e. mathematics), this is fairly straightforward. The rating for the super-criterion should be no higher than the rating for the emphasized criteria. In cases where there are two emphasized criteria (i.e. ELA/Literacy), the overall rating should be no higher than the higher of the two emphasized criteria. The review group will have to consider all of the data in aggregate and make a professional judgment as to whether the ratings of the remaining criteria are enough to pull the rating of the emphasized criteria down.

For example, for Content rating in ELA/Literacy:

- If B.3 and B.5 are Good, the Content rating should be no higher than Good.
- If B.3 is Good and B.5 is Excellent, the Content rating could be Excellent or Good, depending on the ratings of B.6, B.7, and B.8. If they are all Good or Excellent, the rating would be Excellent. If some are Limited, the rating would likely fall to Good.

In all cases, all evidence should be taken into consideration and the decision left to the professional judgment of the review group.

For example, for Depth rating in ELA/Literacy:

- If B.1 and B.2 are Good, the Depth rating should be no higher than Good, even if B.4 and B.9 are Excellent.
- If B.1 is Excellent and B.2 is Good, the Depth rating could be Good or Excellent, depending on the ratings of B.4 and B.9. If they are both Good or Excellent, the rating would be Excellent. If they are both Limited, the rating would likely fall to Good.

In all cases, all evidence should be taken into consideration and the decision left to the professional judgment of the review group.

## Appendix C: Mathematics Scoring Template

### List of Criteria and Sub-Criteria

Criteria & Sub-Criteria	Type
<b>Assesses the <u>content</u> most needed for College and Career Readiness (Cluster)</b>	
<b>Criterion C.1 Focusing strongly on the content most needed for success in later mathematics</b>	
C.1.1 Most important content assessed	Outcome
C.1.2 Assessment design reflect important content	Generalizability
<b>Criterion C.2 Assessing a balance of concepts, procedures, and applications</b>	
C.2.1 Balance of % of points conceptual understanding, procedural skills and fluency, & applications	Outcome
C.2.2 Balance of conceptual understanding, procedural skills and fluency, & applications	Generalizability
C.2.3 Specifications on all math categories for students at all performance levels	Generalizability
<b>Assesses the <u>depth</u> that reflect the demands of College and Career Readiness (Cluster)</b>	
<b>Criterion C.3 Connecting practice to content</b>	
C.3.1 Meaningful connections between practices and content	Outcome
C.3.2. Specifications & explanation of assessing math practices with content	Generalizability
<b>Criterion C.4 Requiring a range of cognitive demand</b>	
C.4.1 Cognitive Demand	Outcome
C.4.2. Specification of Cognitive Demand	Generalizability
<b>Criterion C.5 Ensuring high-quality items and a variety of item types</b>	
C.5.1 Distribution of item types	Outcome
C.5.2 Degree of high-quality items	Outcome
C.5.3. Specification of item quality	Generalizability
C.5.4. Specification of item types	Generalizability

**C.1: Focusing strongly on the content most needed for success in later mathematics:** The assessments help educators keep students on track to readiness by focusing strongly on the content most needed in each grade or course for later mathematics.

	Type	Evidence Descriptors	Location of Evidence	Scoring Guidance	Tentative Cut-Offs
C.1.1	Outcome	<p>The vast majority of score points in each assessment focuses on the content that is most important for students to master in that grade band in order to reach college and career readiness.</p> <p>Goals include:</p> <ul style="list-style-type: none"> <li>In elementary grades, at least three-quarters of the points in each grade align exclusively to the major work of the grade;</li> <li>In middle school grades, at least two-thirds of the points in each grade align exclusively to the major work of the grade; and</li> <li>In high school, at least half of the points in each grade and/or course align exclusively to prerequisites for careers and a wide range of postsecondary studies.</li> </ul> <p>Note: "Major work of the grade" is based on the shifts outlined in the introduction to the CCSS (<a href="http://www.corestandards.org/other-resources/key-shifts-in-mathematics/">http://www.corestandards.org/other-resources/key-shifts-in-mathematics/</a>) and described in the K-8 Publisher's Criteria on page 8 (<a href="http://www.corestandards.org/wp-content/uploads/Math_Publishers_Criteria_K-8_Spring_2013_FINAL1.pdf">http://www.corestandards.org/wp-content/uploads/Math_Publishers_Criteria_K-8_Spring_2013_FINAL1.pdf</a>), which links to</p>	<p>Evidence: Test forms, meta-data</p> <p>Specific metadata from assessment program:</p> <ul style="list-style-type: none"> <li>Point value of item</li> <li>Assigned CCSSM alignment (multiple standards shown, if applicable)</li> </ul> <p>Coding Sheets:</p> <ul style="list-style-type: none"> <li>Do you agree with the assigned alignment? (Y/N)</li> <li>Revised alignment (if needed)</li> <li>Does the item align to Major Work? (N/Major)</li> <li>For High School, does the item align to widely applicable prerequisites? (N/Prerequisite)</li> </ul> <p>Metrics Auto-Calculated:</p> <ul style="list-style-type: none"> <li>Number of items</li> <li>Number and percent of points focused on Major Work.</li> <li>Number and percent of points focused on not-Major Work.</li> <li>Number of aligned items.</li> <li>Percent alignment agreement.</li> <li>Number and percent of Major Work clusters.</li> </ul>	<p>Calculate the percentage of score points that assess the most important content. Assign a score and provide notes under Comments (for each form):</p> <p>For Elementary School:  <b>2 –Meets:</b> At least three-quarters of the score points align exclusively to the Major Work of the grade and all or nearly all Major Work clusters for the grade are assessed.  <b>1 – Partially Meets:</b> At least two-thirds of the score points align exclusively to the Major Work of the grade and the large majority of Major Work clusters for the grade are assessed.  <b>0 – Does Not Meet:</b> Less than two-thirds of the score points align exclusively to the Major Work of the grade and/or less than the majority of the Major Work clusters are assessed.</p> <p>For Middle School:  <b>2 –Meets:</b> At least two-thirds of the score points align exclusively to the Major Work of the grade and all or nearly all Major Work clusters for the grade is assessed.  <b>1 – Partially Meets:</b> More than half of the score points align exclusively to the Major Work of the grade and the large majority of the Major Work clusters for the grade is assessed.  <b>0 – Does Not Meet:</b> Less than half of the score points align exclusively to the Major Work of the grade and/or less than three quarters of the Major Work clusters for the grade are assessed.</p>	<p>For Elementary School:  <b>2 –Meets:</b> 75-100% of score points align exclusively to Major Work and at least 90% of the Major Work clusters are assessed  <b>1 – Partially Meets:</b> 66-74% of the score points align exclusively to Major Work and at least 75% of the Major Work clusters for the grade are assessed  <b>0 – Does Not Meet:</b> 0-65% of the score points align to Major Work and/or less than 75% of the Major Work clusters for the grade are assessed.</p> <p>For Middle School:  <b>2 –Meets:</b> 67-100% of score points align exclusively to the Major Work and at least 90% of the Major Work clusters for the grade are assessed.  <b>1 – Partially Meets:</b> 50-66% of score points align exclusively to the Major Work and at least 75% of the Major Work clusters for the grade are assessed.  <b>0 – Does Not Meet:</b> 0-49% of score points align to the Major Work and/or less than 75% of the Major Work clusters for the grade are assessed.</p>

**C.1: Focusing strongly on the content most needed for success in later mathematics:** The assessments help educators keep students on track to readiness by focusing strongly on the content most needed in each grade or course for later mathematics.

	Type	Evidence Descriptors	Location of Evidence	Scoring Guidance	Tentative Cut-Offs
		<p><a href="http://www.achievethecore.org/downloads/Math%20Shifts%20and%20Major%20Work%20of%20Grade.pdf">http://www.achievethecore.org/downloads/Math%20Shifts%20and%20Major%20Work%20of%20Grade.pdf</a>            showing cluster emphases in footnote 10.</p> <p>"Prerequisites for careers and a wide range of postsecondary studies" are described in the HS Publisher's Criteria on page 8 in Table 1, Criterion #1.            (<a href="http://www.corestandards.org/assets/Math_Publishers_Criteria_HS_Spring%202013_FINAL.pdf">http://www.corestandards.org/assets/Math_Publishers_Criteria_HS_Spring%202013_FINAL.pdf</a>)</p>			

**C.1: Focusing strongly on the content most needed for success in later mathematics:** The assessments help educators keep students on track to readiness by focusing strongly on the content most needed in each grade or course for later mathematics.

	Type	Evidence Descriptors	Location of Evidence	Scoring Guidance	Tentative Cut-Offs
				<p>For High School:</p> <p><b>2 –Meets:</b> At least half of the score points in each course or grade align exclusively to prerequisites for careers and a wide range of postsecondary studies and all or nearly all domains within the widely applicable prerequisites are assessed.</p> <p><b>1 – Partially Meets:</b> Nearly half of the score points in each course or grade align exclusively to prerequisites for careers and a wide range of postsecondary studies and the large majority of domains within the widely applicable prerequisites are assessed.</p> <p><b>0 – Does Not Meet:</b> Less than half of the score points in each course or grade align exclusively to prerequisites for careers and a wide range of postsecondary studies and/or less than the large majority of domains within the widely applicable prerequisites are assessed.</p> <p>Note: For high school end of course assessments, the second part of this scoring guidance regarding domains should be evaluated across the entire set of high school assessments. If only selected end of course assessments are evaluated, each should be evaluated based on the domains relevant to the course.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p>For High School:</p> <p><b>2 –Meets:</b> 50-100% of the score points align exclusively to the widely applicable prerequisites and/or at least 90% of the domains within the widely applicable prerequisites are assessed.</p> <p><b>1 – Partially Meets:</b> 40-50% of the score points align exclusively to the widely applicable prerequisites and at least 75% of the domains are assessed</p> <p><b>0 – Does Not Meet:</b> 0-39% of the score points aligns to the Major Work and/or less than 75% of the domains are assessed.</p> <p>Note: For high school end of course assessments, the second part of this scoring guidance regarding domains should be evaluated across the entire set of high school assessments. If only selected end of course assessments are evaluated, each should be evaluated based on the domains relevant to the course.</p>



**C.1: Focusing strongly on the content most needed for success in later mathematics:** The assessments help educators keep students on track to readiness by focusing strongly on the content most needed in each grade or course for later mathematics.

	Type	Evidence Descriptors	Location of Evidence	Scoring Guidance	Tentative Cut-Offs
C.1.2	Generalizability	<p>The assessment design, including the test blueprints and other specifications, indicate that the vast majority of score points in each assessment focuses on the most important content.</p> <p>Goals include:</p> <ul style="list-style-type: none"> <li>• In elementary grades, at least three-quarters of the points in each grade align exclusively to the Major Work of the grade;</li> <li>• In middle school grades, at least two-thirds of the points in each grade align exclusively to the Major Work of the grade; and</li> <li>• In high school, at least half of the points in each grade and/or course align exclusively to prerequisites for careers and a wide range of postsecondary studies.</li> </ul>	Evidence: Test blueprints and/or other documents identified by the program.	<p>Rate the extent to which the percentage of score points that assess the most important content is indicated in the specifications. Assign a score and provide notes under Comments:</p> <p>For Elementary School:  <b>2 –Meets:</b> The test blueprints or other documents indicate that the large majority of the score points align exclusively to the Major Work of the grade and all or nearly all Major Work clusters for the grade are assessed.  <b>1 – Partially Meets:</b> The test blueprints or other documents indicate that at least two-thirds of the score points align exclusively to the Major Work of the grade and the large majority of Major Work clusters for the grade is assessed.  <b>0 – Does Not Meet:</b> The test blueprints or other documents indicate that less than two-thirds of the score points align exclusively to the Major Work of the grade and/or less than the majority of the Major Work clusters are assessed.</p> <p>For Middle School:  <b>2 –Meets:</b> The test blueprints or other documents indicate that at least two-thirds of the score points align exclusively to the Major Work of the grade and all or nearly all Major Work clusters for the grade is assessed.  <b>1 – Partially Meets:</b> The test blueprints or other documents indicate that more than half of the score points align exclusively to the Major Work of the grade and the large majority of the Major Work clusters for the grade is assessed.  <b>0 – Does Not Meet:</b> The test blueprints or other documents indicate that less than half of the score points align exclusively to the Major Work of the grade and/or less than the majority of the Major Work clusters are assessed.</p>	<p>For Elementary School:  <b>2 –Meets:</b> 75-100% of score points align exclusively to Major Work and at least 90% of the Major Work clusters are assessed  <b>1 – Partially Meets:</b> 66-74% of the score points align exclusively to Major Work and at least 75% of the Major Work clusters for the grade are assessed  <b>0 – Does Not Meet:</b> 0-65% of the score points align to Major Work and/or less than 75% of the Major Work clusters for the grade are assessed.</p> <p>For Middle School:  <b>2 –Meets:</b> 67-100% of score points align exclusively to the Major Work and at least 90% of the Major Work clusters for the grade are assessed.  <b>1 – Partially Meets:</b> 50-66% of score points align exclusively to the Major Work and at least 75% of the Major Work clusters for the grade are assessed.  <b>0 – Does Not Meet:</b> 0-49% of score points align to the Major Work and/or less than 75% of the Major Work clusters for the grade are assessed.</p>

**C.1: Focusing strongly on the content most needed for success in later mathematics:** The assessments help educators keep students on track to readiness by focusing strongly on the content most needed in each grade or course for later mathematics.

	Type	Evidence Descriptors	Location of Evidence	Scoring Guidance	Tentative Cut-Offs
				<p>For High School:</p> <p><b>2 –Meets:</b> The test blueprints or other documents indicate that at least half of score points in each course or grade align exclusively to prerequisites for careers and a wide range of postsecondary studies and all or nearly all domains within the widely applicable prerequisites are assessed.</p> <p><b>1 – Partially Meets:</b> The test blueprints or other documents indicate that nearly half of score points in each course or grade align exclusively to prerequisites for careers and a wide range of postsecondary studies and the large majority of domains within the widely applicable prerequisites are assessed.</p> <p><b>0 – Does Not Meet:</b> The test blueprints or other documents indicate that less than half of score points in each course or grade align exclusively to prerequisites for careers and a wide range of postsecondary studies and/or less than the large majority of the domains within the widely applicable prerequisites are assessed.</p> <p>Note: For high school end of course assessments, the second part of this scoring guidance regarding domains should be evaluated across the entire set of high school assessments. If only selected end of course assessments are evaluated, each should be evaluated based on the domains relevant to the course.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p>For High School:</p> <p><b>2 –Meets:</b> 50-100% of the score points align exclusively to the Major Work and/or less than 75% of the domains within the widely applicable prerequisites are assessed.</p> <p><b>1 – Partially Meets:</b> 40-50% of the score points align exclusively to the Major Work and at least 75% of the domains are assessed</p> <p><b>0 – Does Not Meet:</b> 0-39% of the score points aligns to the Major Work and/or less than 75% of the domains are assessed.</p> <p>Note: For high school end of course assessments, the second part of this scoring guidance regarding domains should be evaluated across the entire set of high school assessments. If only selected end of course assessments are evaluated, each should be evaluated based on the domains relevant to the course.</p>

**C.2: Assessing a balance of concepts, procedures, and applications:** The assessments measure conceptual understanding, fluency and procedural skill, and application of mathematics, as set out in college- and career-ready standards.

	Type	Evidence Descriptors	Location of Evidence	Scoring Guidance	Tentative Cut-Offs
C.2.1	Outcome	<p>The distribution of score points reflects a balance of mathematical concepts, procedures/fluency, and applications.</p> <p>Goals include at least one-quarter of the points come from each of the following categories:</p> <ul style="list-style-type: none"> <li>• Conceptual understanding problems in which students to respond to well-designed conceptual problems;</li> <li>• Procedural skill and fluency problems (e.g., purely procedural problems, some requiring use of efficient algorithms, and others inviting opportunistic strategies); and</li> <li>• Application problems (e.g., in elementary and middle grades, solving grade-appropriate word problems reflecting growing complexity across the grades; in high school, rich application problems requiring students to demonstrate college and career readiness).</li> </ul>	<p>Evidence: Test forms, meta-data</p> <p>Specific metadata from assessment program:</p> <ul style="list-style-type: none"> <li>▪ Point value of item</li> <li>▪ Assigned CCSSM alignment (multiple standards shown, if applicable)</li> </ul> <p>Coding Sheets:</p> <ul style="list-style-type: none"> <li>▪ What does the item assess? <ul style="list-style-type: none"> <li>▪ Conceptual understanding,</li> <li>▪ Procedural skill and fluency,</li> <li>▪ Application,</li> <li>▪ Combined</li> </ul> </li> </ul> <p>Metrics Auto-Calculated:</p> <ul style="list-style-type: none"> <li>▪ Number and percent of points for conceptual understanding, procedural skill and fluency, application, and combined (separate categories).</li> </ul>	<p>Calculate the percentage of score points that assess conceptual understanding, procedural skill and fluency, application, and combined. Assign a score and provide notes under Comments (for each form):</p> <p><b>2 –Meets:</b> At least one quarter and no more than half of the score points are allocated for EACH of the three categories:</p> <ul style="list-style-type: none"> <li>• Conceptual understanding;</li> <li>• Procedural skill and fluency; and</li> <li>• Application.</li> </ul> <p><b>1 – Partially Meets:</b> less than one-quarter of the score points are allocated for one or more of the above three categories.</p> <p><b>0 – Does Not Meet:</b> much less than one-quarter of score points are allocated for one or more of the above three categories.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p><b>2 –Meets:</b> 25-50% are allocated for each of the three categories</p> <p><b>1 – Partially Meets:</b> 19-24% of score points are allocated for one of the three categories</p> <p><b>0 – Does Not Meet:</b> Less than 18% of the score points are allocated for one or more of the three categories</p>

<b>C.2: Assessing a balance of concepts, procedures, and applications:</b> The assessments measure conceptual understanding, fluency and procedural skill, and application of mathematics, as set out in college- and career-ready standards.					
	<b>Type</b>	<b>Evidence Descriptors</b>	<b>Location of Evidence</b>	<b>Scoring Guidance</b>	<b>Tentative Cut-Offs</b>
<b>C.2.2</b>	Generalizability	Test blueprints and other specifications for each grade level specify the distribution of score points, reflecting a balance of mathematical concepts, procedures and fluency, and applications.	Evidence: Test blueprints and/or other documents identified by the program.	<p>Rate the extent to which the test blueprints or other documents reflect a balance of mathematical concepts, procedures/fluency, and applications, as the standards require. Assign a score and provide notes under Comments:</p> <p><b>2 –Meets:</b> The test blueprints or other documents indicate that at least one quarter and no more than half of the score points are allocated for EACH of the three categories:</p> <ul style="list-style-type: none"> <li>• Conceptual understanding;</li> <li>• Procedural skill and fluency; and</li> <li>• Application.</li> </ul> <p><b>1 – Partially Meets:</b> The test blueprints or other documents indicate that less than one-quarter of score points are allocated for one or more of the above three categories.</p> <p><b>0 – Does Not Meet:</b> The test blueprints or other documents indicate that much less than one-quarter of score points are allocated for one or more of the above three categories.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p><b>2 –Meets:</b> 25-50% are allocated for each of the three categories</p> <p><b>1 – Partially Meets:</b> 19-24% of score points are allocated for one of the three categories</p> <p><b>0 – Does Not Meet:</b> Less than 18% of the score points are allocated for one of the three categories</p>

**C.2: Assessing a balance of concepts, procedures, and applications:** The assessments measure conceptual understanding, fluency and procedural skill, and application of mathematics, as set out in college- and career-ready standards.

	Type	Evidence Descriptors	Location of Evidence	Scoring Guidance	Tentative Cut-Offs
<b>C.2.3</b>	Generalizability	Test blueprints and other specifications for each grade level specify that all students, whether high performing or low performing, are required to respond to items within the categories of conceptual understanding, procedural skill and fluency, and applications, so they have the opportunity to show what they know and can do.	Evidence: Test blueprints and/or other documents identified by the program, and /or empirical documentation of distributions of items based on simulations.	<p>Determine the degree of balance of conceptual understanding, procedural skill/fluency, and application for all students regardless of performance level. Assign a score and provide notes under Comments:</p> <p><b>2 –Meets:</b> Documentation indicates that all or nearly all forms balance conceptual understanding, procedural skill and fluency, and application at all performance levels.</p> <p><b>1 – Partially Meets:</b> Documentation indicates that most, but not all, all forms balance conceptual understanding, procedural skill and fluency, and application at all performance levels.</p> <p><b>0 – Does Not Meet:</b> Documentation indicates that many forms will not balance conceptual understanding, procedural skill and fluency, and application.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p><b>Meets:</b> At least 90% of students will be given a form that Meets (score of 2) C.2.2, and the remainder Partially Meet (Score of 1) C.2.2.</p> <p><b>Partially Meets:</b> Fewer than 90% but more than 75% of students will be given a form that Meets C.2.2 OR some students will be given forms that Do Not Meet C.2.2 (score of 0).</p> <p><b>Does Not Meet:</b> Fewer than 75% of students will be given a form that Meets C.2.2 (score of 2)</p>

**C.3: Connecting practice to content:** The assessments include brief questions and also longer questions that connect the most important mathematical content of the grade or course to mathematical practices, for example, modeling and making mathematical arguments.

	Type	Evidence Descriptors	Location of Evidence	Scoring Guidance	Tentative Cut-Offs
C.3.1	Outcome	<p>Assessments for each grade and course meaningfully connect mathematical practices and processes with mathematical content (especially with the most important mathematical content at each grade).</p> <p>Goals include:</p> <ul style="list-style-type: none"> <li>• Every test item that assesses mathematical practices is also aligned to one or more content standards (most often within the Major Work of the grade);</li> <li>• Through the grades, test items reflect growing sophistication of mathematical practices with appropriate expectations at each grade level.</li> </ul>	<p>Evidence: Test forms, meta-data</p> <p>Specific metadata from assessment program:</p> <ul style="list-style-type: none"> <li>▪ Point value of item</li> <li>▪ Assigned CCSSM alignment (multiple standards shown, if applicable)</li> </ul> <p>Coding Sheets:</p> <ul style="list-style-type: none"> <li>▪ If the item measures a mathematical practice, does it align to a content standard? (Y/N)</li> </ul> <p>Metrics Auto-Calculated:</p> <ul style="list-style-type: none"> <li>▪ Number and percent of items measuring practices that also measure content.</li> <li>▪ Number and percent of items measuring practices that do not measure content.</li> </ul>	<p>Calculate the percentage of items that assess mathematical practices and content. Assign a score and provide notes under Comments (for each form):</p> <p><b>2 –Meets:</b> All or nearly all items that assess mathematical practices also align to one or more content standards.</p> <p><b>1 – Partially Meets:</b> The large majority of items that assess mathematical practices also align to one or more content standards.</p> <p><b>0 - Does Not Meet:</b> Less than a large majority of items that assess mathematical practices are aligned to one or more content standards.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p><b>2 –Meets:</b> 90-100% of the items that measure a mathematical practice also align to a content standard.</p> <p><b>1 – Partially Meets:</b> 75-89% of the items that measure a mathematical practice also align to a content standard.</p> <p><b>0 – Does Not Meet:</b> 0-74% of the items that measure a mathematical practice also align to a content standard.</p>

**C.3: Connecting practice to content:** The assessments include brief questions and also longer questions that connect the most important mathematical content of the grade or course to mathematical practices, for example, modeling and making mathematical arguments.

	Type	Evidence Descriptors	Location of Evidence	Scoring Guidance	Tentative Cut-Offs
C.3.2	Generalizability	<p>Item specifications (e.g., task templates, scoring templates) and explanatory materials (e.g. test blueprints and other specifications) specify how mathematical practices will be assessed. Features include meaningful connections for each grade or course between mathematical practices and mathematical content (especially with the most important mathematical content at each grade). Goals include:</p> <ul style="list-style-type: none"> <li>• Every test item that assesses mathematical practices is also aligned to one or more content standards (most often within the Major Work of the grade);</li> <li>• Through the grades, test items reflect growing sophistication of mathematical practices with appropriate expectations at each grade level.</li> </ul>	Evidence: Test blueprints and/or other documents identified by the program.	<p>Assign a score and provide notes under Comments.</p> <p><b>2 –Meets:</b> Documentation indicates that all or nearly all items that assess mathematical practices also align to one or more content standards.</p> <p><b>1 – Partially Meets:</b> Documentation indicates that the large majority of items that assess mathematical practices also align to one or more content standards.</p> <p><b>0 – Does Not Meet:</b> Documentation indicates that less than a large majority of items that assess mathematical practices are aligned to one or more content standards.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p><b>2 –Meets:</b> 90-100%-of the items that measure a mathematical practice also align to a content standard.</p> <p><b>1 – Partially Meets:</b> 75-89% of the items that measure a mathematical practice also align to a content standard.</p> <p><b>0 – Does Not Meet:</b> 0-74% of the items that measure a mathematical practice also align to a content standard.</p>

<b>C.4: Requiring a range of cognitive demand:</b> The assessments require all students to demonstrate a range of higher-order, analytical thinking skills in reading and writing based on the depth and complexity of college- and career-ready standards, allowing robust information to be gathered for students with varied levels of achievement. Assessments include questions, tasks, and prompts about the basic content of the grade or course as well as questions that reflect the complex challenge of college- and career-ready standards.					
	Type	Evidence Descriptors	Location of Evidence	Scoring Guidance	Tentative Cut-Offs
C.4.1	Outcome	<p>The distribution of cognitive demand for each grade level is sufficient to assess the depth and complexity of the state’s standards, as evidenced by use a of generic taxonomy (e.g., Webb’s Depth of Knowledge) or, preferably, classifications specific to the discipline and drawn from mathematical factors, such as</p> <ul style="list-style-type: none"> <li>○ Mathematical topic coverage in the task (single topic vs. two topics vs. three topics vs. four or more topics);</li> <li>○ Nature of reasoning (none, simple, moderate, complex);</li> <li>○ Nature of computation (none, simple numeric, complex numeric or simple symbolic, complex symbolic);</li> <li>○ Nature of application (none, routine word problem, non-routine or less well-posed word problem, fuller coverage of the modeling cycle); and</li> <li>○ Cognitive actions (knowing or remembering, executing, understanding, investigating, or proving).</li> </ul>	<p>Evidence: Test forms Specific metadata from assessment program:</p> <ul style="list-style-type: none"> <li>▪ Point value of item</li> <li>▪ Assigned CCSS alignment (multiple standards shown, if applicable)</li> <li>▪ If program uses Webb, assigned item DoK</li> <li>▪ If program does not use Webb, assigned item cognitive demand level</li> </ul> <p>Coding Sheets:</p> <ul style="list-style-type: none"> <li>▪ By Standard: primary DoK, secondary DoK, tertiary DoK, quaternary DoK.</li> <li>▪ By item: Indicate DoK</li> </ul> <p>Metrics Auto-Calculated: For each test form:</p> <ul style="list-style-type: none"> <li>▪ Number and percent of standards at each of the DoK levels</li> <li>▪ DoK Index = comparing the percentage of score points for items at each DoK level with the percentage of standards at that DoK level, identifying whichever is less, and summing the percentages of the minima</li> <li>▪ DoK Index averaged across both test forms.</li> </ul>	<p>Determine the extent to which the distribution of cognitive demand reflects the cognitive demand of the standards. Assign a score, and provide notes under Comments (for each form).</p> <p><b>2 –Meets:</b> The distribution of cognitive demand of the assessment matches the distribution of cognitive demand of the standards as a whole, AND matches the higher cognitive demand (DoK 3+) of the standards.</p> <p><b>1 – Partially Meets:</b> The distribution of cognitive demand of the assessment partially matches the distribution of cognitive demand of the standards as a whole AND matches the moderate cognitive demand (DoK 2+) of the standards.</p> <p><b>0 – Does Not Meet:</b> The distribution of cognitive demand of the assessment does not match the distribution of cognitive demand of the standards OR has a much higher proportion of low cognitive demand than found in the standards.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p><b>2 – Meets:</b></p> <ul style="list-style-type: none"> <li>• The DoK Index is at least 80% AND</li> <li>• the percentage of score points associated with DoK3+ items is no more than 10% less than the percentage of standards that are DoK3+.</li> </ul> <p><b>1 – Partially Meets:</b></p> <ul style="list-style-type: none"> <li>• The DoK Index is at least 60% AND</li> <li>• the percent of DoK1 score points is no more than 20% higher than the percentage of standards that are DoK1.</li> </ul> <p><b>0 – Does Not Meet:</b></p> <ul style="list-style-type: none"> <li>• The DoK Index is less than 60% OR</li> <li>• the percent of DoK1 score points is more than 20% greater than the percentage of standards that are DoK1.</li> </ul>



<b>C.4: Requiring a range of cognitive demand:</b> The assessments require all students to demonstrate a range of higher-order, analytical thinking skills in reading and writing based on the depth and complexity of college- and career-ready standards, allowing robust information to be gathered for students with varied levels of achievement. Assessments include questions, tasks, and prompts about the basic content of the grade or course as well as questions that reflect the complex challenge of college- and career-ready standards.					
Type	Evidence Descriptors	Location of Evidence	Scoring Guidance	Tentative Cut-Offs	
C.4.2	Generalizability	<p>The distribution of cognitive demand for each grade level is sufficient to assess the depth and complexity of the state's standards, as evidenced by use a of generic taxonomy (e.g., Webb's Depth of Knowledge) or, preferably, classifications specific to the discipline and drawn from mathematical factors, such as</p> <ul style="list-style-type: none"> <li>○ Mathematical topic coverage in the task (single topic vs. two topics vs. three topics vs. four or more topics);</li> <li>○ Nature of reasoning (none, simple, moderate, complex);</li> <li>○ Nature of computation (none, simple numeric, complex numeric or simple symbolic, complex symbolic);</li> <li>○ Nature of application (none, routine word problem, non-routine or less well-posed word problem, fuller coverage of the modeling cycle); and</li> <li>○ Cognitive actions (knowing or remembering, executing, understanding, investigating, or proving).</li> </ul>	<p>Evidence: Test blueprints and/or other documents identified by the program.</p>	<p>Rate the extent to which the documentation specifies that the distribution of cognitive demand reflects the cognitive demand of the standards. . Assign a score and record notes under Comments.</p> <p><b>2 –Meets:</b> Documentation indicates a research-based definition of cognitive demand, a way of operationalizing cognitive demand at the item level, and a rationale for and specification of distribution of cognitive demand for each test form. The distribution of cognitive demand specified matches the distribution of cognitive demand of the standards as a whole. AND matches the higher cognitive demand of the standards.</p> <p><b>1 – Partially Meets:</b> Documentation indicates a definition of cognitive demand, a way of operationalizing cognitive demand at the item level, and a rationale for and specification of distribution of cognitive demand for each test form. The distribution of cognitive demand specified partially matches the distribution of cognitive demand of the standards as a whole AND matches a moderate cognitive demand of the standards.</p> <p><b>0 – Does Not Meet:</b> Documentation does not indicate a definition of cognitive demand, a way of operationalizing cognitive demand at the item level, or a rationale for and specification of distribution of cognitive demand for each test form. The distribution of cognitive demand specified does not match the distribution of cognitive demand of the standards OR does not match the higher or moderate cognitive demands of the standards.</p>	<p><b>2 – Meets:</b></p> <ul style="list-style-type: none"> <li>• If the program uses Webb, the DoK Index is at least 80% AND</li> <li>• the percentage of score points associated with DoK3+ items is no more than 10% less than the percentage of standards that are DoK3+.</li> <li>• If the program uses a measure other than Webb, the definitions, rationales, etc. are appropriate for an assessment program (e.g., specific enough to guide item development and test construction) and the specified distribution of cognitive demand of items on a test form matches the standards as a whole and for the higher demand items/standards.</li> </ul> <p><b>1 – Partially Meets:</b></p> <ul style="list-style-type: none"> <li>• If the program uses Webb, the DoK Index is at least 60% AND</li> <li>• the percent of DoK1 score points is no more than 20% higher than the percentage of standards that are DoK1.</li> </ul>

**C.4: Requiring a range of cognitive demand:** The assessments require all students to demonstrate a range of higher-order, analytical thinking skills in reading and writing based on the depth and complexity of college- and career-ready standards, allowing robust information to be gathered for students with varied levels of achievement. Assessments include questions, tasks, and prompts about the basic content of the grade or course as well as questions that reflect the complex challenge of college- and career-ready standards.

	Type	Evidence Descriptors	Location of Evidence	Scoring Guidance	Tentative Cut-Offs
				<p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the "Location of Evidence" column were not available.</p>	<ul style="list-style-type: none"> <li>• If the program uses a measure other than Webb, the definitions, rationales, etc. are appropriate and the specified distributions of cognitive demand of items on a test form partially matches the standards as a whole and the lower demand items are not significantly disproportional.</li> </ul> <p><b>0 – Does Not Meet:</b></p> <ul style="list-style-type: none"> <li>• If the program uses Webb, the DoK Index is less than 60% OR</li> <li>• the percent of DoK1 score points is more than 20% greater than the percentage of standards that are DoK1.</li> <li>• If the program uses a measure other than Webb, the definitions, rationales, etc. are not appropriate for an assessment program (e.g., too vague to guide item development or test construction) or the specified distribution of cognitive demand of items on a test form does not match that of the standards as a whole or the lower demand items are significantly more than what is in the standards.</li> </ul>

<b>C.5: Ensuring high-quality items and a variety of item types:</b> High-quality items and a variety of item types are strategically used to appropriately assess the standard(s).					
	<b>Type</b>	<b>Evidence Descriptors</b>	<b>Location of Evidence</b>	<b>Scoring Guidelines</b>	<b>Tentative Cut-Offs</b>
<b>C.5.1</b>	Outcome	Items are reviewed to ensure that the distribution of item types for each grade level and content area is sufficient to strategically assess the depth and complexity of the standards being addressed. Item types may include selected-response, short and extended constructed-response, technology-enhanced, and multi-step problems.	<p>Evidence: Test forms, meta-data</p> <p>Specific metadata from assessment program:</p> <ul style="list-style-type: none"> <li>▪ Item type</li> </ul> <p>Coding Sheets:</p> <ul style="list-style-type: none"> <li>• Are there 2 or more item types? (Y/N)</li> <li>• Does at least one of the item types require students to generate, rather than select, a response? (Y/N)</li> </ul> <p>Metrics Auto-Calculated:</p> <ul style="list-style-type: none"> <li>▪ Number and percent of traditional multiple-choice items.</li> <li>▪ Number and percent of multi-select items.</li> <li>▪ Number and percent of evidence-based selected response items.</li> <li>▪ Number and percent of technology enhanced items (does not require student to generate a response).</li> <li>▪ Number and percent of constructed responses.</li> <li>▪ Number and percent of other item type.</li> </ul>	<p>Determine that the distribution of item types is sufficiently used to strategically assess the depth and complexity of the standards being addressed. Assign a score and provide notes under Comments:</p> <p><b>2 –Meets:</b> At least two item formats are used, including one that requires students to generate, rather than select a response (i.e., CR, gridded response).</p> <p><b>1 – Partially Meets:</b> At least two item formats are used but the item formats only require students to select, rather than generate a response.</p> <p><b>0 – Does Not Meet:</b> Only a traditional multiple choice format is used.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p><b>2 –Meets:</b> At least two item formats are used, including one that requires students to generate, rather than select a response (i.e., CR, gridded response).</p> <p><b>1 – Partially Meets:</b> At least two item formats are used but the item formats only require students to select, rather than generate a response.</p> <p><b>0 – Does Not Meet:</b> Only a traditional multiple choice format is used.</p>

<b>C.5: Ensuring high-quality items and a variety of item types:</b> High-quality items and a variety of item types are strategically used to appropriately assess the standard(s).					
	Type	Evidence Descriptors	Location of Evidence	Scoring Guidelines	Tentative Cut-Offs
C.5.2	Outcome	Operational items are reviewed to verify claims of quality, including ensuring the technical quality, alignment to standards, and editorial accuracy of the items	<p>Evidence: Test forms, meta-data</p> <p>Specific metadata from assessment program:</p> <ul style="list-style-type: none"> <li>▪ Point value of item</li> <li>▪ Assigned CCSSM alignment (multiple standards shown, if applicable)</li> <li>▪ Item Type</li> <li>▪ Keyed Correct Answer</li> <li>▪ Rubrics for open-ended items</li> </ul> <p>Coding Sheets:</p> <ul style="list-style-type: none"> <li>▪ Is there a quality issue with this item? (Y/N)</li> <li>▪ If so, what is the issue? (Select all that apply) <ul style="list-style-type: none"> <li>○ Item may not yield valid evidence of targeted skill</li> <li>○ Item has issues with readability</li> <li>○ Item incorrectly keyed</li> <li>○ Item has unintended correct answers</li> <li>○ Mathematically inaccurate</li> </ul> </li> </ul> <p>Metrics Auto-Calculated:</p> <ul style="list-style-type: none"> <li>▪ Number and percent of high-quality items.</li> <li>▪ Number and percent of points by issue type, combined, &amp; total.</li> <li>▪ Number and percent of constructed- and fixed-response types.</li> <li>▪ Number and percent of agreement with given alignment.</li> </ul>	<p>Using the test forms and metadata, determine that there are high-quality items. Assign a score and provide notes under Comments:</p> <p><b>2 –Meets:</b> Nearly all operational items reviewed reflect technical quality, alignment to standards, and editorial accuracy.</p> <p><b>1 – Partially Meets:</b> A few operational items reviewed have issues with technical quality and/or editorial accuracy, and the large majority of items are accurately aligned with the content standards.</p> <p><b>0 – Does Not Meet:</b> Several operational items reviewed have issues with technical quality, alignment to standards, and/or editorial accuracy.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p><b>2 – Meets:</b> 95-100% for editorial and technical; 90% for alignment to standards</p> <p><b>1 – Partially Meets:</b> 90-94% for editorial and technical; 80% for alignment to standards</p> <p><b>0 – Does Not Meet:</b> 0-89% for editorial and technical; 0-79% for alignment to standards</p>

<b>C.5: Ensuring high-quality items and a variety of item types:</b> High-quality items and a variety of item types are strategically used to appropriately assess the standard(s).					
	<b>Type</b>	<b>Evidence Descriptors</b>	<b>Location of Evidence</b>	<b>Scoring Guidelines</b>	<b>Tentative Cut-Offs</b>
<b>C.5.3</b>	Generalizability	<p>To support claims of quality, the following are provided in documentation:</p> <ul style="list-style-type: none"> <li>• Rationales for the use of the specific item types;</li> <li>• Specifications showing the proportion of item types on a form;</li> <li>• For constructed response and performance tasks, a scoring plan (e.g., machine-scored, hand-scored, by whom, how trained), scoring rubrics, and sample student work to confirm the validity of the scoring process;</li> <li>• A description of the process used for ensuring the technical quality, alignment to standards, and editorial accuracy of the items.</li> </ul>	Evidence: Test blueprints, administration and scoring manuals, QC procedure documents, and/or other documents provided by the program.	<p>Assign a score and provide notes under Comments:</p> <p><b>2 –Meets:</b> Documentation supports claims of the technical quality, alignment to standards, and editorial accuracy.</p> <p><b>1 – Partially Meets:</b> Documentation partially supports claims of the technical quality, alignment to standards, and/or editorial accuracy.</p> <p><b>0 – Does Not Meet:</b> Documentation does not support claims of the technical quality, alignment to standards, and/or editorial accuracy.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.</p>	<p><b>2 –Meets:</b> Documentation supports claims of the technical quality, alignment to standards, and editorial accuracy.</p> <p><b>1 – Partially Meets:</b> Documentation partially supports claims of the technical quality, alignment to standards, and/or editorial accuracy.</p> <p><b>0 – Does Not Meet:</b> Documentation does not support claims of the technical quality, alignment to standards, and/or editorial accuracy.</p>

<b>C.5: Ensuring high-quality items and a variety of item types:</b> High-quality items and a variety of item types are strategically used to appropriately assess the standard(s).					
	<b>Type</b>	<b>Evidence Descriptors</b>	<b>Location of Evidence</b>	<b>Scoring Guidelines</b>	<b>Tentative Cut-Offs</b>
<b>C.5.4</b>	Generalizability	Specifications are provided to demonstrate that the distribution of item types for each grade level and content area is sufficient to strategically assess the depth and complexity of the standards being addressed.	Evidence: Test blueprints and/or other documents identified by the program.	<p>Assign a score representing the specification for ensuring high-quality items and a variety of item types; provide notes under Comments:</p> <p><b>2 – Meets:</b> Documentation indicates that at least two item formats should be used, including one that requires students to generate, rather than select, a response (i.e., CR, gridded response).</p> <p><b>1 – Partially Meets:</b> Documentation indicates that at least two formats, but the item formats only require students to select, rather than generate a response.</p> <p><b>0 – Does Not Meet:</b> Documentation indicates that only a traditional multiple choice format is used.</p> <p>Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available</p>	<p><b>2 – Meets:</b> Documentation indicates that at least two item formats should be used, including one that requires students to generate, rather than select, a response (i.e., CR, gridded response).</p> <p><b>1 – Partially Meets:</b> Documentation indicates that at least two formats, but the item formats only require students to select, rather than generate a response.</p> <p><b>0 – Does Not Meet:</b> Documentation indicates that only a traditional multiple choice format is used.</p>

**SCORING SUMMARY**

Criterion		Sub-Criterion	Score		Automatic Criterion-Level Raw Score	Automatic Criterion Score	Group Rating		Automatic Criterion-Level Raw Score	Group Criterion Score Rules
			Form 1	Form 2			Form 1	Form 2		
C.1	Focusing strongly on the content most needed for success in later mathematics	C.1.1			Add (0/1/2) scores from each form and each outcome sub-criterion. Range: 0 to 4	4 = E 3 = G 2 = L 0-1 = W			Add (0/1/2) ratings from each form and each outcome sub-criterion. Range: 0 to 4	E G L W
			☐: Missing	☐: Missing			☐: Missing	☐: Missing		
		Comments:								
		C.1.2			(0/1/2) Rating		Indicate degree of confidence: + : Outcome ratings are likely to be seen in other forms = : Neither confident nor pessimistic - : Outcome ratings are unlikely to be seen in other forms ☒ : Documentation missing			
			☐: Missing							
Comments:										

Criterion	Sub-Criterion	Rating		Automatic Criterion-Level Raw Score	Automatic Criterion Score	Group Rating		Automatic Criterion-Level Raw Score	Group Criterion Score Rules
		Form 1	Form 2			Form 1	Form 2		
C.2	Assessing a balance of concepts, procedures, and applications	C.2.1			Add (0/1/2) ratings from each form and each outcome sub-criterion. Range: 0 to 4			Add (0/1/2) ratings from each form and each outcome sub-criterion. Range: 0 to 4	E G L W
			<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing			<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing	
		Comments:							
		C.2.2			(0/1/2) Rating		Indicate degree of confidence: +: Outcome ratings are likely to be seen in other forms =: Neither confident nor pessimistic -: Outcome ratings are unlikely to be seen in other forms ☒: Documentation missing		
			<input type="checkbox"/> : Missing						
		C.2.3			(0/1/2) Rating				
<input type="checkbox"/> : Missing									
Comments:									



Criterion		Sub-Criterion	Rating		Automatic Criterion-Level Raw Score	Automatic Criterion Score	Group Rating		Automatic Criterion-Level Raw Score	Group Criterion Score Rules	
			Form 1	Form 2			Form 1	Form 2			
C.3	Connecting practice to content	C.3.1			Add (0/1/2) ratings from each form and each outcome sub-criterion. Range: 0 to 4	4 = E 3 = G 2 = L 0-1 = W			Add (0/1/2) ratings from each form and each outcome sub-criterion. Range: 0 to 4	E G L W	
			<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing			<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing			
		Comments:									
		C.3.2				(0/1/2) Rating		Indicate degree of confidence: +: Outcome ratings are likely to be seen in other forms =: Neither confident nor pessimistic -: Outcome ratings are unlikely to be seen in other forms <input checked="" type="checkbox"/> : Documentation missing			
<input type="checkbox"/> : Missing											
Comments:											

Criterion		Sub-Criterion	Rating		Automatic Criterion-Level Raw Score	Automatic Criterion Score	Group Rating		Automatic Criterion-Level Raw Score	Group Criterion Score Rules	
			Form 1	Form 2			Form 1	Form 2			
C.4	Requiring a range of cognitive demand	C.4.1			Add (0/1/2) ratings from each form and each outcome sub-criterion. Range: 0 to 4	4 = E 3 = G 2 = L 0-1 = W			Add (0/1/2) ratings from each form and each outcome sub-criterion. Range: 0 to 4	E G L W	
			☐: Missing	☐: Missing			☐: Missing	☐: Missing			
		Comments:									
		C.4.2			(0/1/2) Rating		Indicate degree of confidence: + : Outcome ratings are likely to be seen in other forms = : Neither confident nor pessimistic - : Outcome ratings are unlikely to be seen in other forms ☒ : Documentation missing				
					☐: Missing						
Comments:											

Criterion		Sub-Criterion	Rating		Automatic Criterion-Level Raw Score	Automatic Criterion Score	Group Rating		Automatic Criterion-Level Raw Score	Group Criterion Score Rules
			Form 1	Form 2			Form 1	Form 2		
C.5	Ensuring high-quality items and a variety of item types	C.5.1	<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing	Add (0/1/2) ratings from each form and each outcome sub-criterion. Range: 0 to 8	7-8= E 5-6 = G 3-4 = L 0-2 = W	<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing	Add (0/1/2) ratings from each form and each outcome sub-criterion. Range: 0 to 8	E G L W
		C.5.2	<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing			<input type="checkbox"/> : Missing	<input type="checkbox"/> : Missing		
		Comments:								
		C.5.3			(0/1/2) Rating		Indicate degree of confidence: += Outcome ratings are likely to be seen in other forms =: Neither confident nor pessimistic -: Outcome ratings are unlikely to be seen in other forms ☒: Documentation missing			
				<input type="checkbox"/> : Missing						
Comments:										

### Cluster Scoring Rules:

The overall rating for the cluster of criteria should not be higher than the rating for the emphasized criteria. In cases where there is one emphasized criterion (i.e., mathematics), this is fairly straightforward. The rating for the cluster should be no higher than the rating for the emphasized criteria. In cases where there are two emphasized criteria (i.e., ELA/Literacy), the overall rating should be no higher than the higher of the two emphasized criteria. The review group will have to consider all of the data in aggregate and make a professional judgment as to whether the ratings of the remaining criteria are enough to pull the rating of the emphasized criteria down.

For example, for Content rating in mathematics (C.1 is the emphasized criterion):

- If C.1 is Good, the Content rating should be no higher than Good, even if C.2 is Excellent.
- If C.1 is Excellent and C.2 is Limited, the Content rating would likely be Good, but could be Excellent.
- In all cases, all evidence should be taken into consideration and the decision is left to the professional judgment of the review group.

For example, for Depth rating in mathematics (C.3 is the emphasized criterion):

- If C.3 is Good, the Depth rating should be no higher than Good, even if C.4 and C.5 are Excellent.
- If C.3 is Good and both C.4 and C.5 are Limited, the Depth rating would likely be Good.
- In all cases, all evidence should be taken into consideration and the decision is left to the professional judgment of the review group.

## Appendix D: Accessibility Scoring Template

### List of Criteria and Sub-Criteria

Criteria & Sub-Criteria	Type
<b>Criterion A.5</b> Providing accessibility to <i>all</i> students, including English learners and students with disabilities (Partial)	
A.5.1.1 Defined the construct, appropriate standardization, and important threats to validity	Generalizability
A.5.1.2 Comprehensive set of coherent procedures	Generalizability
A.5.1.3 Procedures to develop and construct its test forms	Generalizability
A.5.2.1 Appropriate accommodations/access features	Generalizability
A.5.2.2 Appropriate accommodations/access features of Exemplars	Outcome
A.5.3 Validity of accommodations/access features for English learners	Generalizability
A.5.4 Validity of accommodations/access features for students with disabilities	Generalizability

A.5 Providing accessibility to <i>all</i> students, including English learners and students with disabilities (Partial)					
	Type	Evidence Descriptors	Location of Evidence	Scoring Guidance	Tentative Cut-Offs <sup>17</sup>
A.5.1.1	Generalizability	The assessment program has defined the construct, appropriate standardization, and important threats to validity that should be addressed through universal design, accommodations, and access features.	Evidence: Documentation submitted by assessment program (e.g., white papers on defining accessibility for the program that include reviews of the literature, item specifications (including evidence-centered design documents that identify the need for specific accommodations), item review protocols and evidence, empirical evidence from item-tryouts, etc.).	<p><b>2 – Meets:</b> The assessment program has documentation regarding construct definition that is strong and comprehensive, including the following characteristics:</p> <ul style="list-style-type: none"> <li>• defines the construct to be assessed with sufficient clarity that the program and others can distinguish construct-irrelevant from construct-relevant variance;</li> <li>• provides a rationale for the construct definition that incorporates available research;</li> <li>• has defined threats to validity relevant to the assessment program that may require accommodations and/or access features, including those relevant to English learners and students with disabilities;</li> <li>• has a process in place to improve its conception and support of validity regarding accessibility and accommodations.</li> </ul> <p><b>1 – Partially Meets:</b> The assessment program meets at least two but not all of the above characteristics and does not exhibit any of the characteristics of the 0 level.</p> <p><b>0 – Does Not Meet:</b> The assessment program’s documentation manifests one or more of the following characteristics:</p> <ul style="list-style-type: none"> <li>• its definition or rationale is contrary to available research;</li> <li>• its definition and rationale identify the need for specific accommodations/access features but such accommodations/access features are not provided although likely practicable;</li> <li>• meets fewer than two of the characteristics of the 2 level.</li> </ul>	<p>2- The assessment program meets at least 3 of the characteristics for both EL and SWD documentation. It does not meet any of the Level 0 guidance.</p> <p>1- The assessment program meets at least two of the Level 2 characteristics, but not all of them and documentation clearly indicates the program adheres to its policies and procedures regarding accessibility. It does not meet any of the Level 0 guidance.</p> <p>0- Documentation indicates the program meets one or none of the characteristics of Level 2, or documentation indicates the program does not adhere to its development policies or procedures.</p> <p>IE- Use this code if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination.</p>

<sup>17</sup> The final Accessibility Scoring Template published by the Center does not include tentative cut-offs. The cut-offs included in this appendix represent those decided upon by HumRRO and the Center for use in the current study.

A.5 Providing accessibility to <i>all</i> students, including English learners and students with disabilities (Partial)					
	Type	Evidence Descriptors	Location of Evidence	Scoring Guidance	Tentative Cut-Offs <sup>17</sup>
A.5.1.2	Generalizability	The assessment program has a comprehensive set of coherent procedures to develop its items in terms of accessibility, and accommodations receive appropriate attention. The procedures include drawing on research literature, best practice, conceptual analysis, expert review, and empirical data from small-item tryouts (e.g., cognitive labs, focused pilot-testing).	Evidence: Documentation submitted by assessment program (e.g., item specifications (including evidence-centered design documents that identify the need for specific accommodations), item review protocols and evidence, empirical evidence from item-tryouts, etc.).	<p><b>2 – Meets:</b> The assessment program has documentation that is strong and comprehensive regarding development of items with appropriate accessibility, including the following characteristics:</p> <ul style="list-style-type: none"> <li>item development procedures regarding accessibility build on the definitions of the construct established in A.5.1.1 such that accommodations/access features maintain the constructs being assessed and consider the access needs (e.g., cognitive, processing, sensory, physical, language) of the large majority of students;</li> <li>item development procedures regarding accessibility (including instructions for identifying when accommodations/access features may be administered; administration instructions; and scoring instructions) are systematic, e.g., reflecting principles of universal design and sound testing practice, and embodying principles of evidence-centered design or similar practices that make explicit the claims such that they that can be checked conceptually and empirically during design and development that the accommodations/access features reduce construct irrelevant variance (e.g., eliminating unnecessary clutter in graphics, reducing construct-irrelevant reading loads as much as possible)</li> <li>item development procedures include appropriate expert review regarding accessibility at key points in the item development process; the expert review is documented and problems recorded and acted upon; expert review attends to potential challenges due to factors such as disability, ethnicity, culture, geographic location, socioeconomic condition, or gender;</li> <li>item development procedures include appropriate actions based on review of empirical data regarding accessibility at key points in the item development process, such as from cognitive labs or other focused try-outs, pilot-testing, and field-testing. (Analyses based on results from operational administrations will be included in the Test Characteristics evaluation.)</li> </ul>	<p>2- The assessment program meets at least three of the characteristics for EL and SWD documentation.</p> <p>1- The assessment program meets at least two of the Level 2 characteristics, but not all of them. Documentation clearly indicates the program adheres to its policies and procedures regarding accessibility.</p> <p>0- Documentation indicates the program meets one or none of the characteristics of the 2 level, or documentation indicates the program does not adhere to its development policies or procedures.</p> <p>IE- Use this code if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination.</p>

<b>A.5 Providing accessibility to <i>all</i> students, including English learners and students with disabilities (Partial)</b>					
	<b>Type</b>	<b>Evidence Descriptors</b>	<b>Location of Evidence</b>	<b>Scoring Guidance</b>	<b>Tentative Cut-Offs<sup>17</sup></b>
				<p><b>1 – Partially Meets:</b> The assessment program meets at least two but not all of the above characteristics and documentation clearly indicates the program adheres to its policies and procedures regarding accessibility.</p> <p><b>0 – Does Not Meet:</b> Documentation indicates the program meets one or none of the characteristics of the 2 level, or documentation indicates the program does not adhere to its development policies or procedures.</p>	



A.5 Providing accessibility to <i>all</i> students, including English learners and students with disabilities (Partial)					
	Type	Evidence Descriptors	Location of Evidence	Scoring Guidance	Tentative Cut-Offs <sup>17</sup>
A.5.1.3	Generalizability	The assessment program has procedures to develop and construct its test forms while considering accessibility in a way to support valid score inferences.	Evidence: Documentation submitted by assessment program (e.g., white papers on defining accessibility for the program, item specifications (including evidence-centered design documents that identify the need for specific accommodations), item review protocols and evidence, empirical evidence from item-tryouts, etc.).	<p><b>2 – Meets:</b> The assessment program has documentation that is strong and comprehensive regarding development of test forms with appropriate accessibility, including the following characteristics:</p> <ul style="list-style-type: none"> <li>the program has procedures and policies to direct the assembly and administration of test forms for students whose accommodations affect the selection of content of the form (e.g., low vision students who require items that can be appropriately delivered in braille format); the test forms reflect the principles of universal design and sound testing practice;</li> <li>the program has procedures for assigning and delivering the appropriate accommodations/access features to individual students, including assigning special test forms;</li> <li>the program has procedures for detecting and correcting unwanted interactions between multiple accommodations/access features, including accommodations/features offered across multiple items on a form;</li> <li>the program has procedures for collecting, analyzing, and acting on information (including empirical data) to monitor and improve the quality of its test assembly procedures that consider accessibility.</li> </ul> <p><b>1 – Partially Meets:</b> The assessment program meets at least two but not all of the above characteristics and documentation clearly indicates the program adheres to its policies and procedures.</p> <p><b>0 – Does Not Meet:</b> Documentation indicates the program meets one or none of the characteristics of the 2 level, or documentation indicates the program does not adhere to its test form procedures regarding accessibility.</p>	<p>2- The program meets at least three of the characteristics for ELs and SWDs.</p> <p>1- The program meets at least two, but not all of the Level 2 characteristics. Documentation clearly indicates the program adheres to its policies and procedures.</p> <p>0- Documentation indicates the program meets 1 or fewer of the Level 2 characteristics, or documentation indicates the program does not adhere to its test form procedures regarding accessibility.</p> <p>IE- Use this code if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination.</p>

A.5 Providing accessibility to <i>all</i> students, including English learners and students with disabilities (Partial)					
	Type	Evidence Descriptors	Location of Evidence	Scoring Guidance	Tentative Cut-Offs <sup>17</sup>
A.5.2.1	Generalizability	The assessment program offers appropriate accommodations/access features that address the access needs of the large majority of the students intended to be assessed. The available accommodations are documented, including a rationale for how each supports valid score interpretations, when they may be used, and instructions for administration.	Evidence: Documentation submitted by assessment program (e.g., white papers that define construct and appropriate accommodation/accessibility for the program; documents that support the prioritized provision of specific accommodations/access features; documentation supporting the appropriate implementation of the intended accommodations/access features.	<p><b>2 – Meets:</b> The assessment program has documentation that is strong and comprehensive regarding the accommodations/access features the program offers, including:</p> <ul style="list-style-type: none"> <li>• Indication that accommodations/access features are provided by the assessment program for high-moderate incidence needs based on research/data sufficient to support validity of score interpretations, credible use of scores, and legal defensibility, and that no major accessibility needs are unaddressed;</li> <li>• An accurate list of the available accommodations/access features offered by the program, with documentation including relevant construct, rationale, administration/use instructions, scoring instructions (if applicable) (e.g., for magnification, audio representation of graphic elements, linguistic simplification, text-to-speech, speech-to-text, Braille, access to translations and definitions); accommodations are categorized as addressing challenges in presentation, response, setting, and timing and scheduling in test administration;</li> <li>• Information regarding which accommodations/access features are known to be subject to variations in administration frequency due to policy (e.g., required/prohibited/permissible by a state or other user group), and technical information on possible impact on validity and comparability of score interpretations due to such policy variations. (Empirical information welcome here, but optional; will be required in Test Characteristics evaluation.);</li> <li>• If it is reasonably expected that there will be variation, then there is a clear policy regarding differentiating scores of students who have variations that change the construct sufficiently to invalidate the scores, including not combining those scores with those of the bulk of students when computing or reporting scores.</li> </ul> <p><b>1 – Partially Meets:</b> The assessment program meets the first bullet and at least three additional bullets but not all of the above characteristics and documentation clearly indicates the program adheres to its policies and procedures regarding accessibility.</p>	<p>2 – The program completely meets all of four of the characteristics for ELs and SWDs.</p> <p>1- The assessment program meets the first Level 2 bullet and at least 3 additional Level 2 bullets but not all of the above characteristics and documentation clearly indicates the program adheres to its policies and procedures regarding accessibility.</p> <p>0- Documentation indicates the program does not meet the first Level 2 bullet, or meets 3 or fewer of the other Level 2 characteristics, or documentation indicates the program does not adhere to its policies and procedures regarding accessibility.</p> <p>IE- Use this code if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination.</p>

A.5 Providing accessibility to <i>all</i> students, including English learners and students with disabilities (Partial)					
	Type	Evidence Descriptors	Location of Evidence	Scoring Guidance	Tentative Cut-Offs <sup>17</sup>
				<p><b>0 – Does Not Meet:</b> Documentation indicates the program does not meet the first bullet, or meets fewer than three of the other characteristics of the 2 level, or documentation indicates the program does <i>not</i> adhere to its policies and procedures regarding accessibility.</p>	
A.5.2.2	Outcomes	<p>The assessment program offers appropriate accommodations/access features that address the access needs of the large majority of the students intended to be assessed. The available accommodations are documented, including a rationale for how each supports valid score interpretations, when they may be used, and instructions for administration.</p>	<p>10-25 Exemplars of accommodations/access features, of which at least 5 will be in conjunction with the most widely used accommodations/access features in the program.</p> <p>An Exemplar may be an assessment item with a highlighted accommodation; an Exemplar may be a tool that may be applied to many items (e.g., a tool that the student may use to highlight text on instructions or reading passages); an Exemplar may illustrate some aspect of accessibility in the instructions, navigation design, or other general design of the assessment (e.g., the use of plain language, clear visual design, etc.). Each Exemplar will have accompanying documentation that annotates the construct the Exemplar is intended to assess, what the accommodation/access feature is, how it supports more valid score interpretations, instructions for administration, and validity evidence.</p>	<p><b>2 – Meets:</b> The Accessibility Exemplars and accompanying documentation provided by the assessment program indicate adequate coverage of major access/accommodations needs with acceptable quality for all or almost all of the Exemplars. Acceptable quality includes construct focus and ease of use.</p> <p><b>1 – Partially Meets:</b> The Accessibility Exemplars and accompanying document provided by the assessment program indicates either adequate coverage of major access/accommodations needs OR acceptable quality for the Exemplars provided.</p> <p><b>0 – Does Not Meet:</b> The Accessibility Exemplars and accompanying documentation provided by the assessment program indicates neither adequate coverage of major access/accommodations needs nor adequate quality.</p>	<p>2- The Accessibility Exemplars and accompanying documentation provided by The assessment program indicate adequate coverage of major access/accommodations needs with acceptable quality for all or almost all of The Exemplars. Acceptable quality includes construct focus and ease of use.</p> <p>1- The Accessibility Exemplars and accompanying document provided by the assessment program indicates either adequate coverage of major access/accommodations needs OR acceptable quality for the Exemplars provided.</p> <p>0- The Accessibility Exemplars and accompanying documentation provided by the assessment program indicates neither adequate coverage of major access/accommodations needs nor adequate quality.</p> <p>IE- Use this code if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination.</p>

A.5 Providing accessibility to <i>all</i> students, including English learners and students with disabilities (Partial)					
	Type	Evidence Descriptors	Location of Evidence	Scoring Guidance	Tentative Cut-Offs <sup>17</sup>
A.5.3	Generalizability	The program's consideration of validity and available accommodations/access features specifically address the needs of students who are English learners.	Evidence: Documentation submitted by assessment program (e.g., white papers on defining accessibility for the program that include reviews of the literature, item specifications (including evidence-centered design documents that identify the need for specific accommodations), item review protocols and evidence, empirical evidence from item-tryouts, etc.).	<p><b>2 – Meets:</b> Documentation indicates the assessment program “Meets” both A.5.1 (parts A.5.1.1, 5.1.2, and 5.1.3) and A.5.2 (parts A.5.2.1 and 5.2.2) regarding English learners.</p> <p><b>1 – Partially Meets:</b> Documentation indicates the assessment program at least “Partially Meets” both A.5.1 (parts A.5.1.1, 5.1.2, and 5.1.3) and A.5.2 (parts A.5.2.1 and 5.2.2) for English learners, but does not “Meet” both regarding English learners.</p> <p><b>0 – Does Not Meet:</b> Documentation indicates the program “Does Not Meet” at least A.5.1 or A.5.2 regarding English learners.</p>	<p>2- Documentation indicates the assessment program “Meets” both A.5.1 (parts A.5.1.1, 5.1.2, and 5.1.3) and A.5.2 (parts A.5.2.1 and 5.2.2) regarding English learners.</p> <p>1- Documentation indicates the assessment program at least “Partially Meets” both A.5.1 (parts A.5.1.1, 5.1.2, and 5.1.3) and A.5.2 (parts A.5.2.1 and 5.2.2) for English learners, but does not “Meet” both regarding English learners.</p> <p>0– Documentation indicates the program “Does Not Meet” at least A.5.1 or A.5.2 regarding English learners.</p> <p>IE- Use this code if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination.</p>
A.5.4	Generalizability	The program's consideration of validity and available accommodations/access features specifically address the needs of students with disabilities.	Evidence: Documentation submitted by assessment program (e.g., white papers on defining accessibility for the program that include reviews of the literature, item specifications (including evidence-centered design documents that identify the need for specific accommodations), item review protocols and evidence, empirical evidence from item-tryouts, etc.).	<p><b>2 – Meets:</b> Documentation indicates the assessment program “Meets” both A.5.1 (parts A.5.1.1, 5.1.2, and 5.1.3) and A.5.2 (parts A.5.2.1 and 5.2.2) regarding students with disabilities.</p> <p><b>1 – Partially Meets:</b> Documentation indicates the assessment program at least “Partially Meets” both A.5.1 (parts A.5.1.1, 5.1.2, and 5.1.3) and A.5.2 (parts A.5.2.1 and 5.2.2) for students with disabilities, but does not “Meet” both regarding students with disabilities.</p> <p><b>0 – Does Not Meet:</b> Documentation indicates the program “Does Not Meet” at least A.5.1 or A.5.2 regarding students with disabilities.</p>	<p>2- Documentation indicates the assessment program “Meets” both A.5.1 (parts A.5.1.1, 5.1.2, and 5.1.3) and A.5.2 (parts A.5.2.1 and 5.2.2) regarding SWDs.</p> <p>1- Documentation indicates the assessment program at least “Partially Meets” both A.5.1 (parts A.5.1.1, 5.1.2, and 5.1.3) and A.5.2 (parts A.5.2.1 and 5.2.2) for English</p>

<b>A.5 Providing accessibility to <i>all</i> students, including English learners and students with disabilities (Partial)</b>					
	<b>Type</b>	<b>Evidence Descriptors</b>	<b>Location of Evidence</b>	<b>Scoring Guidance</b>	<b>Tentative Cut-Offs<sup>17</sup></b>
					<p>learners, but does not “Meet” both regarding SWDs.</p> <p>0– Documentation indicates the program “Does Not Meet” at least A.5.1 or A.5.2 regarding SWDs.</p> <p>IE- Use this code if there is insufficient information to score. Comments must be added to explain rationale for insufficient information determination.</p>

**SCORING SUMMARY**

Criterion		Sub-Criterion	Score	Automatic Criterion-Level Raw Score	Automatic Criterion Score	Group Criterion Score Rules
A.5.1	Following the principles of universal design	A.5.1.1		Add (0/1/2) scores from A.5.1.1, A.5.1.2, A.5.1.3 & A.5.2.1. Range: 0 to 8	7-8 = E 5-6 = G 3-4 = L 0-2 = W	E G L W
			<input type="checkbox"/> : Missing			
		Comment:				
		A.5.1.2				
			<input type="checkbox"/> : Missing			
		Comment:				
		A.5.1.3				
	<input type="checkbox"/> : Missing					
	Comment:					
A.5.2	Offering appropriate accommodations/access features	A.5.2.1				
			<input type="checkbox"/> : Missing			
		Comment:				
A.5.2	Offering appropriate accommodations/access features	A.5.2.2	(0/1/2 Score)	Indicate degree of confidence: +: Exemplars helped reduce interference of measuring the focal construct. Exemplars appear to be clear and easy to use. =: Neither helped nor distracted -: Exemplars did not help reduce interference of measuring the focal construct. Exemplars were not clear and easy to use. ☒: Documentation missing		
			<input type="checkbox"/> : Missing			
		Comment:				
A.5.3	English learners	A.5.3				
			<input type="checkbox"/> : Missing			
		Comment:				
A.5.4	Students with disabilities	A.5.4				
			<input type="checkbox"/> : Missing			
		Comment:				

## Appendix E: Metadata for Test Content Evaluation Methodology

### **ELA/Literacy:**

#### Passages

1. Passage identifier
2. Type of passage/text type (e.g., informational)
3. Passage grade designation
4. Passage permissioned or commissioned designation
5. Passage cognitive demand
6. Quantitative text complexity
7. Qualitative text complexity

#### Items

8. Item/entity ID
9. Item position
10. Item grade designation
11. Operational/field test item
12. Maximum possible score points for each item
13. Scoring rubrics for multi-point items
14. Assigned CCSS alignment
15. Item type
16. Keyed correct answer
17. Item cognitive demand
18. For items with stimuli, type of media

### **Mathematics:**

1. Item/entity ID
2. Item position
3. Item grade designation
4. Operational/field test item
5. Maximum possible score points for each item
6. Scoring rubrics for multi-point items
7. Assigned CCSS alignment
8. Mathematical practice designation
9. Cognitive demand taxonomy
10. Item cognitive demand
11. Item type
12. Keyed correct answer
13. For items with stimuli, type of media

## Appendix F: Reviewer Biographies

### *Reviewer Biographies: Outcome*

**Mary Blaker** is from Parkersburg, West Virginia, has a master's degree, and has been teaching high school English language arts for 10 years. Ms. Blaker currently teaches courses in English and advanced communications to students in grades 9–10. She has experience teaching students with disabilities and English language learners. Ms. Blaker participated in a Smarter Balanced alignment study and served as a member of the Smarter Balanced Digital Library State Network of Educators.

**Tiffany Clapsaddle** is from Murphy, North Carolina and has 18 years teaching experience. Ms. Clapsaddle holds an education specialist degree in curriculum and instruction. She currently serves as mathematics director for accountability and curriculum. Within the past 3 years, she has taught courses in mathematics II and Advanced Placement statistics and has experience teaching students with disabilities and English language learners. Ms. Clapsaddle has been a curriculum director, item writer for the North Carolina Department of Public Instruction, participant in a Smarter Balanced alignment study, and a mathematics partner on Project LEAD.

**Victoria Collaro** is from Reno, Nevada and has 23 years teaching experience. Ms. Collaro has a master's degree in mathematics and currently serves as program coordinator for grades 6–12 mathematics. As a classroom teacher, she had experience teaching students with disabilities and English language learners. Ms. Collaro has provided Common Core State Standards implementation training to mathematics teachers of students in grades 6–12, facilitated others in writing district course guides, and led vendor product reviews for supplemental materials. Ms. Collaro is a Core Advocate for the Student Achievement Partners and she has presented at the National Council of Teachers of Mathematics.

**Thomas Coy** is from Little Rock, Arkansas and has 11 years teaching experience. Mr. Coy received an undergraduate minor degree in special education and a master's degree in mathematics education. He currently serves as a Secondary Mathematics Specialist for Curriculum and Assessment at the Arkansas Department of Education. Mr. Coy has experience teaching both students with disabilities and English language learners. He was a member of the rapid response feedback team for the standards writing group, facilitated a statewide review of the standards for Arkansas, and he helped write the PARCC model content frameworks.

**Rebecca Curtright** is from Reno, Nevada and has 10 years teaching experience. Over the past 3 years, Ms. Curtright has taught algebra I and algebra II honors courses to students in grades 9–12. As a classroom teacher, she has experience teaching students with disabilities and English language learners. Ms. Curtright is on special assignment working in the Assessment Department where she is responsible for writing, implementing, and collecting data related to instructional materials and district Common Mathematics Finals for algebra I, geometry, algebra II, and algebra honors. Her experiences include writing items for the district's common mathematics assessments, participating in state item review panels, serving on vendor product review committees, and developing Nevada's new state assessments aimed at vetting their alignment to the standards. Ms. Curtright also leads professional development and collaboration



sessions for educators in implementing the standards within the district. Ms. Curtright has a master's degree in teacher leadership.

**Heather Goodwin-Nelson** is from Orem, Utah and has 15 years teaching experience. As a classroom teacher, she has experience with English language learners. Over the past 3 years, she has taught special education mathematics and English language arts classes to students in kindergarten through grade 12. Ms. Goodwin-Nelson currently works at the Utah Virtual Academy with special education English language arts teachers as an instructional coach and special education facilitator. She received her bachelor's degree in Elementary Education from Brigham Young University-Hawaii in 1994. During her undergraduate work, Ms. Goodwin-Nelson taught various grade levels before receiving a master's degree in Special Education from Brigham Young University

**Jessica Hunter** is from Monroe, Louisiana and has 6 years teaching experience. She currently teaches high school geometry and, as a classroom teacher, she has experience teaching students with disabilities. Ms. Hunter has created instructional reviews for the Louisiana Department of Education. Ms. Hunter has a master's degree in mathematics education and is currently working on her doctoral degree.

**Elizabeth Keatley** is from Delbarton, West Virginia and has 19 years teaching experience. Over the past 3 years, Ms. Keatley has taught courses in English, speech, and journalism to students in grades 9–12. As a classroom teacher, she has experience teaching students with disabilities and English language learners. Ms. Keatley participated in a Smarter Balanced alignment study. Ms. Keatley is a member of West Virginia's State Network of Educators, which includes making submissions to the Smarter Balanced Library.

**Tim LaVan** is from Shippen, Pennsylvania and has 25 year teaching experience. He holds a bachelor's degree in mathematics education and a master's degree in computational science, and completed doctoral work in in curriculum and instruction. Over the past 3 years, Mr. LaVan has taught courses in algebra I, algebra II, discrete mathematics, algebra lab, and calculus. As a classroom teacher, he has experience teaching students with disabilities and English language learners. Mr. LaVan is a member of Pennsylvania's Mathematics Keystone Exam Advisory Committee. He also served as a member of Pennsylvania's committee for PARCC.

**Marissa McClish** is from Reno, Nevada and has almost 7 years teaching experience. She currently serves as a mathematics trainer for teachers of students in grades 6–12, providing professional development for the region. As a high school classroom teacher, she taught algebra (regular, honors, remedial), geometry, advanced algebra, trigonometry, mathematics analysis, integrated science, and Earth science. She has experience teaching students with disabilities and English language learners; she is Cross-cultural, Language, and Academic Development (CLAD) certified in California. Ms. McClish is a Student Achievement Partners (SAP) Core Advocate and presented at the first national SAP Advocates Summit in 2015. She also works with six counties in Nevada, where she has trained K–12 teachers in alignment of assessments to the rigor called for in the Common Core State Standards. Ms. McClish holds a master's of education in curriculum and instruction (mathematics and science emphasis).

**John Neal** is from Alexandria, Louisiana and has 16 years teaching experience. Mr. Neal has a master's degree in education and currently serves as an instructional coach. As a classroom teacher, he taught English IV and Advanced Placement literature to students in grade 12. He has experience teaching students with disabilities and English learners. Mr. Neal's familiarity and experience with the Common Core State Standards include creating lessons that adapt special education students' individual needs yet are coherent to the standards, reviewing curriculum materials for their alignment to the standards, and serving on item review committees. Additionally, as a member of the Smarter Balanced Digital Library State Network of Educators, Mr. Neal has reviewed various ELA/literacy and mathematics lessons and tasks for alignment to the standards.

**Susan Newton** is from Camp Hill, Pennsylvania and has over 30 years teaching experience. As a classroom teacher, she has taught honors algebra I, college preparatory and honors pre-calculus, and college preparatory and honors algebra II to students in grades 9–12. She has experience teaching students with disabilities and English language learners. Ms. Newton's experience includes serving as an item reviewer for Pennsylvania and a member of Pennsylvania's Mathematics Keystone Exam Advisory Committee. She holds a master's degree in mathematics education.

**Lacey Noel** is from Lafayette, Louisiana and has 5 years teaching experience. Over the past 3 years, Ms. Noel taught English II courses to students in grade 10. As a classroom, she has experience teaching students with disabilities and English language learners. Ms. Noel has a master's degree in educational leadership and is currently an instructional strategist, working with high school teachers in all content areas to implement the Common Core State Standards, align curriculum to the standards, and ensure fidelity of teaching practices. Ms. Noel's familiarity and experience with the Common Core State Standards include providing professional development seminars to the district's teachers on implementing the standards and ensuring assessments align to the standards. Ms. Noel has worked with Student Achievement Partners as a Core Advocate and is a member of the Literacy Delivery Team in Louisiana.

**Tabitha Pacheco** is from Springville, Utah and has 8 years teaching experience. She is a National Board-certified teacher in exceptional needs, a 2015 National Teaching Fellow for the Hope Street Group, and serves on the Practitioners Advisory Group for The Centers on Great Teachers and Leaders. For the past nine years, Ms. Pacheco has worked with students with disabilities and is an expert on accommodating and scaffolding the CCSS to meet the needs of all learners. She has a bachelor's degree from Brigham Young University in Family Life and a post-baccalaureate degree in special education from Brigham Young University.

**Samantha Singer Swafford** is from Nashville, Tennessee, has a bachelor's degree in secondary English education, and has taught high school for 5 years. Ms. Singer Swafford currently teaches courses in English language arts, reading, English, and Advanced Placement literature to students in grades 7 and 10–12. As a classroom teacher, Ms. Singer Swafford has experience teaching students with disabilities and English language learners. Her experience includes facilitating workshops to train high school ELA/literacy teachers in the metropolitan Nashville Public Schools about implementing the standards. Ms. Singer Swafford is currently part of the team that is reevaluating the Common Core State Standards Scope and Sequence documents for high school.

**Rachel Saunders** is from Southbury, Connecticut and has 10 years teaching experience. She currently serves as a mathematics instructional coach in the Danbury Public Schools. As a classroom teacher, Ms. Saunders taught mathematics, pre-algebra, algebra I, geometry, and algebra II to students in grades 6–11, including students with disabilities and English language learners. Ms. Saunders' familiarity and experience with the mathematics Common Core State Standards includes writing curriculum for the Danbury school district at both middle and high school levels; working for LearnZillion.com creating online videos, lessons, and coaching others; and serving as a participant in two Smarter Balanced alignment studies. Ms. Saunders received her master's degree in mathematics education.

**Rachel Snell** is from Pinconning, Michigan and has 14 years teaching experience. Ms. Snell has a master's degree and is currently serving as a secondary assessment and instructional coach. As a classroom teacher, Ms. Snell has taught classes in English, speech, and journalism to students in grades 9–12. She also has taught civics, reading, U.S. history, world history, and psychology. Her higher education teaching includes pre-teaching Common Core State Standards classes at Saginaw Valley State University. Her familiarity and experience with the Common Core State Standards also includes participating in a Smarter Balanced alignment study, and designing assessments and professional development related to the ELA/literacy Common Core State Standards. She has experience teaching students with disabilities.

**Diana Walker** is from Reno, Nevada and has 22 years teaching experience. Dr. Walker earned her doctoral degree and currently serves as a K–12 literacy learning facilitator, with expertise in 3–16 English language arts and K–12 English language arts. Over the past 3 years, Dr. Walker has developing professional development materials designed to support English language arts teachers in learning about and implementing the Common Core State Standards. As a classroom teacher, Dr. Walker had experience teaching students with disabilities and she currently teaches English language arts endorsement courses for teachers.

**Charlie Wayne** is from Shamokin, Pennsylvania and has 5 years teaching experience at elementary and middle schools, post-secondary institutes, and for business. He has been an assessment specialist in mathematics with the Pennsylvania Department of Education (PDE) for over 17 years. Mr. Wayne has been Pennsylvania's mathematics representative with PARCC (Pennsylvania is a participating state in PARCC and Smarter Balanced). He has also participated in various alignment studies with Dr. Norman Webb, Achieve, the American Association for the Advancement of Science (AAAS), and HumRRO. Mr. Wayne has a bachelor's degree in economics and a master's degree in mathematics along with a graduate certificate in large-scaled assessment education. Prior to coming to PDE, he taught in elementary and middle schools, at the post-secondary level, and at a business institute.

**Henry Wyborney** is from Cheney, Washington, has a bachelor's degree, and has been a high school teacher for over 28 years. Mr. Wyborney has taught courses in English language arts, reading, and Advanced Placement literature to students in grades 7 and 10–12. As a classroom teacher, Mr. Wyborney has experience teaching students with disabilities and English language learners. His experiences include participating on item development committees, pilot testing assessments, serving as item line descriptor writer, and participating on a cut-score committee.

### **Reviewer Biographies: Generalizability**

**Heather Goodwin-Nelson** is from Orem, Utah and has 15 years teaching experience. As a classroom teacher, she has experience with English language learners. Over the past 3 years, she has taught special education mathematics and English language arts classes to students in kindergarten through grade 12. Ms. Goodwin-Nelson currently works at the Utah Virtual Academy with special education English language arts teachers as an instructional coach and special education facilitator. She received her bachelor's degree in Elementary Education from Brigham Young University-Hawaii in 1994. During her undergraduate work, Ms. Goodwin-Nelson taught various grade levels before receiving a master's degree in Special Education from Brigham Young University

**Tim LaVan** is from Shippen, Pennsylvania and has 25 year teaching experience. He holds a bachelor's degree in mathematics education and a master's degree in computational science, and completed doctoral work in curriculum and instruction. Over the past 3 years, Mr. LaVan has taught courses in algebra I, algebra II, discrete mathematics, algebra lab, and calculus. As a classroom teacher, he has experience teaching students with disabilities and English language learners. Mr. LaVan is a member of Pennsylvania's Mathematics Keystone Exam Advisory Committee. He also served as a member of Pennsylvania's committee for PARCC.

**Tabitha Pacheco** is from Springville, Utah and has 8 years teaching experience. She is a National Board-certified teacher in exceptional needs, a 2015 National Teaching Fellow for the Hope Street Group, and serves on the Practitioners Advisory Group for The Centers on Great Teachers and Leaders. For the past nine years, Ms. Pacheco has worked with students with disabilities and is an expert on accommodating and scaffolding the CCSS to meet the needs of all learners. She has a bachelor's degree from Brigham Young University in Family Life and a post-baccalaureate degree in special education from Brigham Young University.

**Charlie Wayne** is from Shamokin, Pennsylvania and has 5 years teaching experience at elementary and middle schools, post-secondary institutes, and for business. He has been an assessment specialist in mathematics with the Pennsylvania Department of Education (PDE) for over 17 years. Mr. Wayne has been Pennsylvania's mathematics representative with PARCC (Pennsylvania is a participating state in PARCC and Smarter Balanced). He has also participated in various alignment studies with Dr. Norman Webb, Achieve, the American Association for the Advancement of Science (AAAS), and HumRRO. Mr. Wayne has a bachelor's degree in economics and a master's degree in mathematics along with a graduate certificate in large-scaled assessment education. Prior to coming to PDE, he taught in elementary and middle schools, at the post-secondary level, and at a business institute.

### **Reviewer Biographies: Accessibility**

**Jamal Abedi** is a Professor of educational measurement at the University of California, Davis. His research interests include psychometrics and test development. His recent works include studies on the validity of assessment, accommodation, and classification for English language learners (ELLs) and ELLs with disabilities. Dr. Abedi serves on assessment advisory boards for a number of states and assessment consortia as an expert in testing for ELLs. He is the recipient of the 2003 *Outstanding Contribution Relating Research to Practice* award by the

American Educational Research Association, *the 2008 Lifetime Achievement Award* by the California Educational Research Association, the 2013 National Association of Test Directors: *Outstanding Contribution to Educational Assessment* and the 2014 University of California, Davis: *Distinguished Scholarly Public Service Award*. He holds a Master's degree in psychology and a PhD degree in psychometrics from Vanderbilt University.

**Daniel Anderson** is a Research Associate for Behavioral Research and Teaching at the University of Oregon. He earned his PhD in Educational Research Methodology from the University of Oregon in 2015, with an emphasis in measurement. Dr. Anderson currently serves as the lead psychometrician for the Alternate Assessment for students with significant cognitive disabilities for the state of Oregon, and was formerly the project manager of a large federal grant funded to develop a classroom-based assessment with built-in features of universal design.

**Laurene Christensen** is a Research Associate at the National Center on Educational Outcomes (NCEO). In this position, she works with states to improve outcomes for students with disabilities and English language learners, particularly in the area of assessment accommodations. Recent projects at NCEO involve analyzing emerging issues from the federal standards and assessments peer review. Dr. Christensen has expertise in large-scale assessments, school accountability, language acquisition, research design, and transition/postsecondary issues. She is the author of a number of publications on accommodations for students with disabilities, and has also written and published materials on assessment issues for English language learners, both with and without disabilities. Dr. Christensen previously served as a consultant to both PARCC and Smarter Balanced as they developed their accommodations frameworks.

**Heather Goodwin-Nelson** is from Orem, Utah and has 15 years teaching experience. As a classroom teacher, she has experience with English language learners. Over the past 3 years, she has taught special education mathematics and English language arts classes to students in kindergarten through grade 12. Ms. Goodwin-Nelson currently works at the Utah Virtual Academy with special education English language arts teachers as an instructional coach and special education facilitator. She received her bachelor's degree in Elementary Education from Brigham Young University-Hawaii in 1994. During her undergraduate work, Ms. Goodwin-Nelson taught various grade levels before receiving a master's degree in Special Education from Brigham Young University.

**Audrey Lesondak** is an Education Consultant at the Wisconsin Department of Public Instruction (WI DPI), where she managed assessments for English learners in the Office of Student Assessment, and now coordinates Title III initiatives. Prior to her work at WI DPI, Ms. Lesondak worked with diverse populations in Chicago, advocating in the areas of housing and homelessness, and then served for a decade as a teacher of English learners. Her assessment-related work at WI DPI encompassed providing state review and workgroup support for English learner translations and accommodations in the new, online consortia-developed tests. Ms. Lesondak received her BA with a concentration in German and her MA in Urban Planning and Policy from the University of Illinois at Chicago, as well as her post-Baccalaureate teaching licensure from Concordia University. She was a Fulbright-Hays travel abroad fellow, and has served as the board president for the Wisconsin Teachers of Speakers of Other Languages (WITESOL).

**Vitaliy Shyyan** is a Research Associate at the National Center on Educational Outcomes where he works with state departments of education to improve outcomes for diverse students, including students with disabilities, English language learners, and English language learners with disabilities. His duties include overseeing the Center's leadership and coordination efforts; conducting research and evaluation that inform the improvement of accountability assessments for states and consortia; collaborating with the Center's personnel on publications, products, tools, and services; and designing and delivering technical assistance to states and assessment consortia. Dr. Shyyan also has expertise in large-scale assessments, accessibility and accommodations, research and evaluation design, language acquisition, and intercultural education. He reviewed PARCC's Accessibility Features and Accommodations Manual and co-wrote Smarter Balanced's Usability, Accessibility, and Accommodations Guidelines.

**Lynn Shafer Willner** is an ELL Accessibility Researcher at the World-Class Instructional Design and Assessment (WIDA) at the Wisconsin Center for Educational Research (WCER) where she supports WIDA's development of research, materials, and guidance for educators who work with ELLs with disabilities. Dr. Shafer Willner is a member of the WIDA Assessment and Standards teams and works closely with the Professional Learning team. Previously, she worked at WestEd, supporting the development and implementation of state and national English Language Development standards (serving as lead author of the ELPA21 English Language Proficiency Standards) and conducting Common Core State Standards alignment studies. She also worked at The George Washington University Center for Equity and Excellence in Education where she helped SEAs refine their ELL accommodation policies and guideline and created online trainings to support their implementation. She has a Ph.D. from George Mason University in education, a master's from the University at Buffalo in elementary education, and a bachelor's in history and political science from the University of Rochester. Dr. Shafer Willner also helped develop the first draft of the ELL section of PARCC's Accommodation Manual and contributed to the research and development of the initial draft of the Smarter Balanced Usability, Accessibility, and Accommodations Guidelines.

**Sara (Bolt) Witmer** is an associate professor of school psychology at Michigan State University and a Nationally Certified School Psychologist. Her research focuses on examining assessment tools that can enhance instructional decision-making for students who are at-risk for poor academic outcomes. She also conducts research on accommodations for diverse learners (e.g., students with disabilities, English language learners), and more generally on methods for the effective inclusion of all students in large-scale assessment and accountability programs.

**Joy Zabala** is Director of Technical Assistance for CAST and the Co-Director of its federally funded National Center on Accessible Educational Materials for Learning (AEM Center). She was previously the Director of Technical Assistance for the Accessible Instructional Materials (AIM) Consortium (2007-2009) and the National Center on Accessible Instructional Materials (2009-2014). Dr. Zabala is a leading expert on the use of assistive technology (AT) to improve education and living for people with disabilities. As a technologist, special educator, teacher trainer, and conference speaker, she has earned international recognition for her work on AT and Universal Design for Learning (UDL). Dr. Zabala has also been involved in the review of both PARCC and Smarter Balanced assessment materials and accommodations as a part of her work at CAST and the National AEM Center.

## Appendix G: ACT Aspire Criteria B and C Ratings and Summary Statements

### ACT ASPIRE – HIGH SCHOOL ENGLISH LANGUAGE ARTS/LITERACY SUMMATIVE ASSESSMENT

#### OVERALL SUMMARY

ACT Aspire received a “Weak Match” for **Content** on its early high school ELA/literacy summative assessment. Reviewers found that less than the recommended large majority of items on this assessment required close reading and analysis of text. Additionally, most items did not focus on central ideas and important particulars, and less than the majority was aligned to the specifics of the standards. Although the items required students to refer to the text to find an answer, the majority of items did not require students to support their answers citing evidence from the text; the criteria recommends that more than half the score points be based on items requiring direct use of textual evidence. For programs that do not include narrative writing (such as ACT Aspire), the criteria recommend that expository and argumentative writing types be represented across forms in the grade band. However, reviewers found that this assessment included only a single expository writing prompt. The large majority of items that assessed writing standards was multiple-choice and did not require students to actually generate a written response. Additionally, neither writing prompt required students to confront text or other stimuli directly, to draw on textual evidence, or to support valid inferences from text, as recommended by the criteria. Across forms, reviewers found that very few tier 2 words (that is, words commonly used in written texts, often referred to as “general academic words”) were used to assess vocabulary. Although the majority of items required students to use context to determine meaning, most did not assess words important to central ideas, as recommended by the criteria. Per the criteria, vocabulary and language skills should be reported as sub-scores or at least 13% of score points should be devoted to assessing each skill. Although language skills were reported as a sub-score, vocabulary was not nor were sufficient score points devoted to assessing vocabulary. Many items involved editing, which mirrors real world activity, as recommended by the criteria. Another area in which ACT Aspire met the criteria is that the large majority of items that assessed research and inquiry required students to analyze, synthesize, organize, and use information.

ACT Aspire received a “Good Match” for **Depth** on its early high school ELA/literacy summative assessment. As recommended by the criteria, approximately two-thirds of the texts were informational and nearly all of the passages were previously published or of publishable quality. The majority of informational passages were expository rather than narrative in structure; however, the passages were not split evenly for literary nonfiction, history/social science, and science/technical (most of the passages were history/social science). Per the criteria, quantitative and qualitative measures should be used to place each text at the appropriate grade band and level. As recommended by the criteria, reviewers found that the distribution of cognitive demand of the assessment matched the distribution of cognitive demand of the standards as a whole. Additionally, reviewers found that the percentage of score points associated with DOK levels 3 and 4 approximately matched the percentage of standards at DOK levels 3 and 4. For both forms, reviewers found that at least two item formats were used and that one of those formats required students to generate a response, as recommended by the criteria. In terms of item quality, reviewers believed the assessment lacked technical quality because of the poor alignment to the stated grade-level standards (the criteria recommends that at least 90% of items align). Reviewers also felt items had readability issues because students were not provided specific instructions for responding to the various item types.

## ACT ASPIRE – HIGH SCHOOL ENGLISH LANGUAGE ARTS/LITERACY SUMMATIVE ASSESSMENT<sup>a</sup>






### I. CONTENT: Assesses the content most needed for College and Career Readiness

Reviewers found that less than the recommended large majority of items on this assessment required close reading and analysis of text. Additionally, most items did not focus on central ideas and important particulars, and less than the majority was aligned to the specifics of the standards. Although the items required students to refer to the text to find an answer, the majority of items did not require students to support their answers citing evidence from the text; the criteria recommends that more than half the score points be based on items requiring direct use of textual evidence. For programs that do not include narrative writing (such as ACT Aspire), the criteria recommends that expository and argumentative writing types be represented across forms in the grade band. However, reviewers found that this assessment included only a single expository writing prompt. The large majority of items that assessed writing standards was multiple-choice and did not require students to actually generate a written response. Additionally, neither writing prompt required students to confront text or other stimuli directly, to draw on textual evidence, or to support valid inferences from text, as recommended by the criteria. Across forms, reviewers found that very few tier 2 words (that is, words commonly used in written texts, often referred to as "general academic words") were used to assess vocabulary. Although the majority of items required students to use context to determine meaning, most did not assess words important to central ideas, as recommended by the criteria. Per the criteria, vocabulary and language skills should be reported as sub-scores or at least 13% of score points should be devoted to assessing each skill. Although language skills were reported as a sub-score, vocabulary was not nor were sufficient score points devoted to assessing vocabulary. Many items involved editing, which mirrors real world activity, as recommended by the criteria. Another area in which ACT Aspire met the criteria is that the large majority of items that assessed research and inquiry required students to analyze, synthesize, organize, and use information.

Criteria	Rating	Group Summary Statement
<b>B.3 Reading.</b> <sup>b</sup> Require students to read closely and use specific evidence from texts to obtain and defend correct responses.		Reviewers found that less than the recommended large majority of items on this assessment required close reading and analysis of text. Additionally, most items did not focus on central ideas and important particulars, and less than the majority was aligned to the specifics of the standards. Although the items required students to refer to the text to find an answer, the majority of score points did not require students to support their answers citing evidence from the text. The criteria recommend that more than half the score points be based on items requiring direct use of textual evidence.
<b>B.5 Writing.</b> Require students to engage in close reading and analysis of texts. Across grade band, tests include balance of expository, persuasive/argument, and narrative writing.		For programs that do not include narrative writing (such as ACT Aspire), the criteria recommend that expository and argumentative writing types be represented across forms in the grade band. However, reviewers found that each form included only a single expository writing prompt. The large majority of items that assessed writing standards was multiple-choice and did not require students to actually generate a written response. Additionally, neither writing prompt required students to confront text or other stimuli directly, to draw on textual evidence, or to support valid inferences from text, as recommended by the criteria.  Note: All items that were aligned to a writing standard were included in the evaluation of Criterion B.5, regardless of whether the item required students to actually generate a written response.



ACT ASPIRE – HIGH SCHOOL ENGLISH LANGUAGE ARTS/LITERACY SUMMATIVE ASSESSMENT		
Criteria	Rating	Group Summary Statement
B.6 Vocabulary and language skills. Place sufficient emphasis on academic vocabulary and language conventions used in real-world activities.		Reviewers found that very few tier 2 words (that is, words commonly used in written texts, often referred to as "general academic words") were used to assess vocabulary. Although the majority of items required students to use context to determine meaning, most did not assess words important to central ideas, as recommended by the criteria. Many items involved editing, which mirrors a real world activity, as recommended by the criteria. Per the criteria, vocabulary and language skills should be reported as sub-scores or at least 13% of score points should be devoted to assessing each skill. Although language skills were reported as a sub-score, vocabulary was not nor were sufficient score points devoted to assessing vocabulary.
B.7 Research and inquiry. Require students to demonstrate the ability to find, process, synthesize and organize information from multiple sources.		As recommended by the criteria, reviewers found that the large majority of items that assessed research and inquiry required students to analyze, synthesize, organize, and use information.
B.8 Speaking and listening. <sup>c</sup> Over time and as advances allow, measure speaking and listening skills.		None of the items assessed speaking and listening skills required for college and career readiness; the criteria recommend assessing Speaking and Listening skills over time and as advances allow. Thus, this criterion was not included when establishing the composite Content rating (indicated by gray shading).

<sup>a</sup> Legend: E = Excellent, G = Good, L = Limited, W = Weak, IE = Insufficient Evidence

<sup>b</sup> Underlined criteria indicate criteria that are to have more weight in the composite rating.

<sup>c</sup> Cells that have gray shading were not considered when establishing the composite rating.



## ACT ASPIRE – HIGH SCHOOL ENGLISH LANGUAGE ARTS/LITERACY SUMMATIVE ASSESSMENT<sup>a</sup>

G

### II. DEPTH: Assesses the depth that reflect the demands of College and Career Readiness

As recommended by the criteria, approximately two-thirds of the texts were informational and nearly all of the passages were previously published or of publishable quality. The majority of informational passages were expository rather than narrative in structure; however, the passages were not split evenly for literary nonfiction, history/social science, and science/technical (most of the passages were history/social science). As recommended by the criteria, reviewers found that the distribution of cognitive demand of the assessment matched the distribution of cognitive demand of the standards as a whole. Additionally, as recommended by the criteria, reviewers found that the percentage of score points associated with DOK levels 3 and 4 approximately matched the percentage of standards at DOK levels 3 and 4. For both forms, reviewers found that at least two item formats were used and that one of those formats required students to generate a response, as recommended by the criteria. Reviewers believed that the assessment lacked technical quality because of the poor alignment to the grade-level standards (the criteria requires that at least 90% of items align). Reviewers also felt items had readability issues because students were not provided specific instructions for responding to the various item types.

Criteria	Rating	Group Summary Statement
<p><b><u>B.1 Text quality and types.</u></b><sup>b</sup> Include aligned balance of high-quality literary and informational texts.</p>	<div style="border: 1px solid #0056b3; border-radius: 50%; width: 30px; height: 30px; display: flex; align-items: center; justify-content: center; margin: 0 auto;">G</div>	<p>As recommended by the criteria, approximately two-thirds of the texts were informational and nearly all of the passages were previously published or of publishable quality. The majority of informational passages were expository rather than narrative in structure; however, the passages were not split evenly for literary nonfiction, history/social science, and science/technical (most of the passages were history/social science), as recommended by the criteria.</p> <p>It should be noted there are typically a limited number of passages that can be included on any given assessment, thus, the methodology’s recommended distribution of passage types could be influenced greatly by a single discrepancy that might result in a different (lower or higher) rating.</p>
<p><b><u>B.2 Complexity of texts.</u></b><sup>c</sup> Passages are at appropriate levels of text complexity, increasing through the grades, and multiple forms of authentic, high-quality texts are used.</p>	<div style="border: 1px solid #0056b3; border-radius: 50%; width: 30px; height: 30px; display: flex; align-items: center; justify-content: center; margin: 0 auto;">G</div>	<p>Per the criteria, quantitative and qualitative measures should be used to place each text at the appropriate grade band and level. The ACT Aspire program documentation indicated the use of both quantitative and qualitative measures of text complexity; however, reviewers could not provide a rating based on the items because it was not possible to obtain complexity metadata for all programs that participated in this study in a format for the reviewers to evaluate.</p> <p>Note: The Criterion B.2 rating is based solely on program documentation as reviewers were not able to rate the extent to which quantitative measures are used to place each text in a grade band. Thus, reviewers did not consider the Criterion B.2 rating when establishing the composite Depth rating (indicated by the gray shading).</p>

ACT ASPIRE – HIGH SCHOOL ENGLISH LANGUAGE ARTS/LITERACY SUMMATIVE ASSESSMENT		
Criteria	Rating	Group Summary Statement
B.4 Cognitive demand. <sup>d</sup> Distribution of cognitive demand for each grade level is sufficient to assess the depth and complexity of the standards.		As recommended by the criteria, reviewers found that the distribution of cognitive demand of the assessment matched the distribution of cognitive demand of the standards as a whole. Additionally, as recommended by the criteria, reviewers found that the percentage of score points associated with DOK levels 3 and 4 approximately matched the percentage of standards at DOK levels 3 and 4.  It should be noted that assessment programs implement different cognitive complexity frameworks that might impact this rating, especially those programs that do not use Webb’s DOK methodology to determine cognitive demand, as used in this study. ACT Aspire uses Webb’s DOK methodology.
B.9 High-quality items and a variety of item types. <sup>e</sup> Items are of high technical and editorial quality and each test form includes at least two item types including at least one that requires students to generate a response.		Reviewers found that at least two item formats were used and that one of those formats required students to generate a response, as recommended by the criteria. In terms of item quality, reviewers believed that the assessment lacked technical quality because of the poor alignment to the stated grade-level standards (the criteria require that at least 90% of items align). <sup>3</sup> Reviewers also felt items had readability issues because students were not provided specific instructions for responding to the various item types.

<sup>a</sup> Legend: E = Excellent, G = Good, L = Limited, W = Weak, IE = Insufficient Evidence

<sup>b</sup> Underlined criteria indicate criteria that are to have more weight in the composite rating.

<sup>c</sup> Cells that have gray shading were not considered when establishing the composite rating.

<sup>d</sup> The DOK distribution of the grade 11-12 standards were used for all four assessment programs for comparison purposes; research by WestEd found the DOK distribution of the grade 9-10 standards was not substantively different from the DOK distribution of the grade 11-12 standards (WestEd. (2011). Smarter Balanced Assessment Consortium Common Core State Standards Analysis: Eligible Content for the Summative Assessment. Prepared for the Smarter Balanced Assessment Consortium: Edynn Sato, Rachel Lagunoff, and Peter Worth).

<sup>e</sup>The ACT Aspire items included in this study were aligned to the Common Core State Standards (CCSS) College and Career Readiness Anchor Standards rather than to grade-level standards, as recommended by the CCSSO criteria. Although requested, ACT did not provide the grade-level standard alignment; therefore, HumRRO staff identified the most appropriate grade-level standard(s) for each ACT Aspire ELA/literacy and mathematics item based on the aligned Anchor Standards. The grade-level standards alignment identified by HumRRO staff was used in this study.

## ACT ASPIRE – HIGH SCHOOL MATHEMATICS SUMMATIVE ASSESSMENT

### OVERALL SUMMARY

ACT Aspire received a “Limited Match” for **Content** on its early high school mathematics summative assessment. Reviewers found that many of the widely applicable prerequisites assessed were below the high school level. Additionally, fewer than half the score points were aligned to the high school level widely applicable prerequisites, which is fewer than that recommended by the criteria. In addition, many domains/standards in the widely applicable prerequisites were not assessed. The criteria recommend that at least one-quarter to half the score points be allocated each to conceptual understanding, procedural skill and fluency, and application; however, reviewers found that the forms reviewed for this study varied widely in their distribution of score points allocated to each of the three categories (across forms, reviewers found a very low percentage of items assessed application).

ACT Aspire received a “Good Match” for **Depth** on its early high school mathematics assessment. As recommended by the criteria, all items that assessed a Mathematical Practice also aligned to at least one content standard. Although the criteria recommend the distribution of cognitive demand of the assessment match the distribution of cognitive demand of the standards, reviewers found that the distribution of cognitive demand of this assessment only partially matched the cognitive demand of the standards. Specifically, reviewers found that this assessment included a lower percentage of score points at DOK level 2 and a higher percentage of score points at DOK levels 1 and 3 than were expected by the standards. Per the criteria, this assessment included at least two item formats and one of those formats required students to generate rather than select a response. However, reviewers found that many items did not align well to the standards and that many of the items aligned to off-grade level standards. Additionally, reviewers identified an excessive reading load for a number of items.

## ACT ASPIRE – HIGH SCHOOL MATHEMATICS SUMMATIVE ASSESSMENT<sup>a</sup>

### I. CONTENT: Assesses the content most needed for College and Career Readiness L

Reviewers found that many of the widely applicable prerequisites assessed across the two forms were below the high school level. Additionally, fewer than half the score points were aligned to the high school level widely applicable prerequisites, which is fewer than that recommended by the criteria. The criteria also recommend that all or nearly all domains/standards within the widely applicable prerequisites be assessed; however, reviewers found there were many domains/standards in the widely applicable prerequisites that were not assessed. The criteria recommend that at least one-quarter to half the score points be allocated each to conceptual understanding, procedural skill and fluency, and application; however, reviewers found that the forms reviewed for this study varied widely in their distribution of score points allocated to each of the three categories (reviewers found a very low percentage of items assessed application).

Criteria	Rating	Group Summary Statement
<u>C.1 Focus.</u> <sup>b</sup> Tests focus strongly on content most needed in each grade or course for success in later mathematics (prerequisites for careers and a wide range of postsecondary studies).	<span style="background-color: #ffc107; border-radius: 50%; padding: 10px; font-weight: bold; color: white; font-size: 24px;">L</span>	Reviewers found that many of the widely applicable prerequisites assessed across the two forms were below the high school level. Additionally, fewer than half the score points were aligned to the high school level widely applicable prerequisites, which is fewer than that recommended by the criteria. In addition, many domains/standards in the widely applicable prerequisites were not assessed.
C.2 Concepts, procedures, and applications. Place balanced emphasis on measurement of conceptual understanding, fluency and procedural skill, and application of mathematics.	<span style="background-color: #dc3545; border-radius: 50%; padding: 10px; font-weight: bold; color: white; font-size: 24px;">W</span>	The criteria recommend that at least one-quarter to half the score points be allocated each to conceptual understanding, procedural skill and fluency, and application; however, reviewers found that both forms varied widely in their distribution of score points allocated to each of the three categories (across forms, reviewers found a very low percentage of items assessed application).

<sup>a</sup> Legend: E = Excellent, G = Good, L = Limited, W = Weak, IE = Insufficient Evidence




<sup>b</sup> Underlined criteria indicate criteria that are to have more weight in the composite rating.

## ACT ASPIRE – HIGH SCHOOL MATHEMATICS SUMMATIVE ASSESSMENT<sup>a</sup>

G

### II. DEPTH: Assesses the depth that reflect the demands of College and Career Readiness

As recommended by the criteria, all items that assessed a Mathematical Practice also aligned to at least one content standard. Although the criteria recommend the distribution of cognitive demand of the assessment match the distribution of cognitive demand of the standards, reviewers found that the distribution of cognitive demand of this assessment only partially matched the cognitive demand of the standards. Specifically, reviewers found that this assessment included a lower percentage of score points at DOK level 2 and a higher percentage of score points at DOK levels 1 and 3 than were expected by the standards. Per the criteria, this assessment included at least two item formats and one of those formats required students to generate rather than select a response. However, reviewers found that many items did not align well to the standards and that many of the items aligned to off-grade level standards. Additionally, reviewers identified an excessive reading load for a number of items.

Criteria	Rating	Group Summary Statement
<b><u>C.3 Connecting practice to content.</u></b> <sup>b</sup> Questions meaningfully connect mathematical practices and processes with mathematical content.		As recommended by the criteria, all items that assessed a Mathematical Practice also aligned to at least one content standard.
C.4 Cognitive demand. Distribution of cognitive demand for each grade level is sufficient to assess the depth and complexity of the standards.		Although the criteria recommend the distribution of cognitive demand of the assessment match the distribution of cognitive demand of the standards, reviewers found that the distribution of cognitive demand of this assessment only partially matched the cognitive demand of the standards. Specifically, reviewers found this assessment included a lower percentage of score points at DOK level 2 and a higher percentage of score points at DOK levels 1 and 3 than were expected by the standards.  It should be noted that assessment programs implement different cognitive complexity frameworks that might impact this rating, especially those programs that do not use Webb’s DOK methodology to determine cognitive demand, as used in this study. ACT Aspire uses Webb’s DOK methodology.
C.5 High-quality items and a variety of item types. Items are of high technical and editorial quality and each test form includes at least two item types including at least one that requires students to generate a response.		Per the criteria, this assessment included at least two item formats and one of those formats required students to generate rather than select a response. However, reviewers found that many items did not align well to the standards and that many of the items aligned to off-grade level standards. Additionally, reviewers identified an excessive reading load for a number of items.

<sup>a</sup> Legend: E = Excellent, G = Good, L = Limited, W = Weak, IE = Insufficient Evidence

<sup>b</sup> Underlined criteria indicate criteria that are to have more weight in the composite rating.

## Appendix H: MCAS Criteria B and C Ratings and Summary Statements

### MCAS – HIGH SCHOOL ENGLISH LANGUAGE ARTS/LITERACY SUMMATIVE ASSESSMENT

#### OVERALL SUMMARY

MCAS received a “Limited Match” for **Content** on its high school ELA/literacy summative assessment. The large majority of items required close reading. Nearly all of the items focused on central ideas. Although many of the items referenced the text, not all of the responses required direct textual evidence. Although the writing prompt on this assessment intended students to provide an expository response, it was written in such a way that students would provide a narrative response. The prompt required students to write about a previously read passage but it did not require the response to cite direct textual evidence. Further, the prompt asked students to retell a series of events without making inferences or fully comprehending any concept. The large majority of items that assessed vocabulary used tier 2 words (that is, words commonly used in written texts, often referred to as “general academic words”); however, not all items required students to reference the text for context. Less than half of the items that assessed language skills mirrored real-world activities (as compared to the large majority recommended by the criteria). Per the criteria, language and vocabulary skills should be reported as sub-scores or at least 13% of score points should be devoted to these skills; while language skills were reported as a sub-score, vocabulary was not nor was there an adequate percentage of score points devoted to assessing vocabulary. None of the test items assessed research and inquiry that mirrored real-world activities, so none of the items required analysis, synthesis, or organization of research information.

MCAS received a “Limited Match” for **Depth** on its high school ELA/literacy summative assessment. This assessment was judged to have the appropriate levels of text complexity; however, less than two-thirds of the passages were informational, as recommended by the criteria. Slightly more than half of the informational passages were expository in nature (that is, writing that explains or informs about a specific topic) rather than virtually all as recommended by the criteria. Additionally, only two of the three reading types were addressed (literary nonfiction, history/social/science, science/technical) while the criteria requires a balance among the three writing types. The criteria recommend that nearly all passages be previously published or of publishable quality; although the passages were previously published, reviewers did not find that they represented a wide range of text structures and purposes. Many items required a lower level of cognitive demand compared to what was required in the standards. Reviewers found the distribution of cognitive demand of this assessment only partially matched the distribution of cognitive demand of the standards as a whole; there was too much coverage of the lower levels of cognitive demand and many questions did not require a high level of strategic or extended thinking. Per the criteria, at least two item types were included on this assessment and one of those item types required students to generate a response. Additionally, the items reflected technical quality and editorial accuracy; however, the representation of and alignment to the standards could have been better.

## MCAS – HIGH SCHOOL ENGLISH LANGUAGE ARTS/LITERACY SUMMATIVE ASSESSMENT<sup>a</sup>

### I. CONTENT: Assesses the content most needed for College and Career Readiness



L

The large majority of items required close reading. Nearly all of the items focused on central ideas. Although many of the items referenced the text, not all of the responses required direct textual evidence. Although the writing prompt on this assessment intended students to provide an expository response, it was written in such a way that students would provide a narrative response. The prompt required students to write about a previously read passage but it did not require the response to cite direct textual evidence. Further, the prompt asked students to retell a series of events without making inferences or fully comprehending any concept. The large majority of items that assessed vocabulary used tier 2 words (that is, words commonly used in written texts, often referred to as “general academic words”); however, not all items required students to reference the text for context. Less than half of the items that assessed language skills mirrored real-world activities (as compared to the large majority recommended by the criteria). Per the criteria, language and vocabulary skills should be reported as sub-scores or at least 13% of score points should be devoted to these skills; while language skills were reported as a sub-score, vocabulary was not nor was an adequate percentage of score points devoted to assessing vocabulary. None of the test items assessed research and inquiry that mirrored real-world activities, so none of the items required analysis, synthesis, or organization of research information.

Criteria	Rating	Group Summary Statement
<b>B.3 Reading.</b> <sup>b</sup> Require students to read closely and use specific evidence from texts to obtain and defend correct responses.	<span style="background-color: #90ee90; border-radius: 50%; width: 30px; height: 30px; display: flex; align-items: center; justify-content: center; margin: 0 auto;">G</span>	The large majority of items required close reading. Nearly all of the items focused on central ideas. Although many of the items referenced the text, not all of the responses required direct textual evidence.
<b>B.5 Writing.</b> Require students to engage in close reading and analysis of texts. Across grade band, tests include balance of expository, persuasive/argument, and narrative writing.	<span style="background-color: #ff0000; border-radius: 50%; width: 30px; height: 30px; display: flex; align-items: center; justify-content: center; margin: 0 auto;">W</span>	<p>Although the writing prompt on this assessment intended students to provide an expository response, it was written in such a way that students would provide a narrative response. The prompt required students to write about a previously read passage but it did not require the response to cite direct textual evidence. Further, the prompt asked students to retell a series of events without making inferences or fully comprehending any concept.</p> <p>Note: All items that were aligned to a writing standard were included in the evaluation of Criterion B.5, regardless of whether the item required students to actually generate a written response.</p>
B.6 Vocabulary and language skills. Place sufficient emphasis on academic vocabulary and language conventions used in real-world activities.	<span style="background-color: #ffcc00; border-radius: 50%; width: 30px; height: 30px; display: flex; align-items: center; justify-content: center; margin: 0 auto;">L</span>	The large majority of items that assessed vocabulary used tier 2 words (that is, words commonly used in written texts, often referred to as “general academic words”); however, not all items required students to reference the text for context. Less than half of the items that assessed language skills mirrored real-world activities (as compared to the large majority recommended by the criteria). Per the criteria, language and vocabulary skills should be reported as sub-scores or at least 13% of score points should be devoted to these skills; while language skills were reported as a sub-score, vocabulary was not nor was there an adequate percentage of score points devoted to assessing vocabulary.



## MCAS – HIGH SCHOOL ENGLISH LANGUAGE ARTS/LITERACY SUMMATIVE ASSESSMENT

Criteria	Rating	Group Summary Statement
B.7 Research and inquiry. Require students to demonstrate the ability to find, process, synthesize and organize information from multiple sources.		None of the test items assessed research and inquiry that mirrored real-world activities, so none of the items required analysis, synthesis, or organization of research information.
B.8 Speaking and listening. <sup>c</sup> Over time and as advances allow, measure speaking and listening skills.		None of the items assessed speaking and listening skills required for college and career readiness; the criteria recommend assessing Speaking and Listening skills over time and as advances allow. Thus, this criterion was not included when establishing the composite Content rating (indicated by gray shading).

<sup>a</sup> Legend: E = Excellent, G = Good, L = Limited, W = Weak, IE = Insufficient Evidence

<sup>b</sup> Underlined criteria indicate criteria that are to have more weight in the composite rating.

<sup>c</sup> Cells that have gray shading were not considered when establishing the composite rating.

## MCAS – HIGH SCHOOL ENGLISH LANGUAGE ARTS/LITERACY SUMMATIVE ASSESSMENT<sup>a</sup>



### II. DEPTH: Assesses the depth that reflect the demands of College and Career Readiness

L

This assessment was judged to have the appropriate levels of text complexity; however, less than two-thirds of the passages were informational, as recommended by the criteria. Slightly more than half of the informational passages were expository in nature (that is, writing that explains or informs about a specific topic) rather than virtually as recommended by the criteria. Additionally, only two of the three reading types were addressed (literary nonfiction, history/social/science, science/technical) while the criteria recommends a balance among the three writing types. The criteria recommend that nearly all passages be previously published or of publishable quality; although the passages were previously published, reviewers did not find that they represented a wide range of text structures and purposes. Many items required a lower level of cognitive demand compared to what was required in the standards. Reviewers found the distribution of cognitive demand of this assessment only partially matched the distribution of cognitive demand of the standards as a whole; there was too much coverage of the lower levels of cognitive demand and many questions did not require a high level of strategic or extended thinking. Per the criteria, at least two item types were included on this assessment and one of those item types required students to generate a response. Additionally, the items reflected technical quality and editorial accuracy; however, the representation of and alignment to the standards could have been better.

Criteria	Rating	Group Summary Statement
<p><b><u>B.1 Text quality and types.</u></b><sup>b</sup> Include aligned balance of high-quality literary and informational texts.</p>	G	<p>This assessment was judged to have the appropriate levels of text complexity; however, less than two-thirds of the passages were informational, as recommended by the criteria. Slightly more than half of the informational passages were expository in nature (that is, writing that explains or informs about a specific topic) rather than virtually all as recommended by the criteria. Additionally, only two of the three reading types were addressed (literary nonfiction, history/social/science, science/technical) while the criteria recommend a balance among the three writing types. The criteria recommend that nearly all passages be previously published or of publishable quality; although the passages were previously published, reviewers did not find that they represented a wide range of text structures and purposes.</p> <p>It should be noted there are typically only a limited number of passages that can be included on any given assessment, thus, the methodology's recommended distribution of passage types could be influenced greatly by a single discrepancy that might result in a different (lower or higher) rating.</p>
<p><b><u>B.2 Complexity of texts.</u></b><sup>c</sup> Passages are at appropriate levels of text complexity, increasing through the grades, and multiple forms of authentic, high-quality texts are used.</p>	G	<p>Per the criteria, quantitative and qualitative measures should be used to place each text at the appropriate grade band and level. The MCAS program documentation indicated the use of both quantitative and qualitative measures of text complexity; however, reviewers could not provide a rating based on the items because it was not possible to obtain complexity metadata from all programs included in this study in a format for the reviewers to evaluate.</p> <p>Note: The Criterion B.2 rating is based solely on program documentation as reviewers were not able to rate the extent to which quantitative measures are used to place each text in a grade band. Thus, reviewers did not consider the Criterion B.2 rating when establishing the composite Depth rating (indicated by the gray shading).</p>

## MCAS – HIGH SCHOOL ENGLISH LANGUAGE ARTS/LITERACY SUMMATIVE ASSESSMENT

Criteria	Rating	Group Summary Statement
B.4 Cognitive demand. <sup>d</sup> Distribution of cognitive demand for each grade level is sufficient to assess the depth and complexity of the standards.		<p>Many items required a lower level of cognitive demand compared to what was required in the standards. Reviewers found the distribution of cognitive demand of this assessment only partially matched the distribution of cognitive demand of the standards as a whole; there was too much coverage of the lower levels of cognitive demand and many questions did not require a high level of strategic or extended thinking.</p> <p>It should be noted that assessment programs implement different cognitive complexity frameworks that might impact this rating, especially those programs that do not use Webb’s DOK methodology to determine cognitive demand, as used in this study. MCAS uses the National Assessment of Educational Progress (NAEP) model for cognitive complexity.</p>
B.9 High-quality items and a variety of item types. Items are of high technical and editorial quality and each test form includes at least two item types including at least one that requires students to generate a response.		<p>Per the criteria, at least two item types were included on this assessment and one of those item types required students to generate a response. Additionally, the items reflected technical quality and editorial accuracy; however, the representation of and alignment to the standards could have been better.</p>

<sup>a</sup> Legend: E = Excellent, G = Good, L = Limited, W = Weak, IE = Insufficient Evidence

<sup>b</sup> Underlined criteria indicate criteria that are to have more weight in the composite rating.

<sup>c</sup> Cells that have gray shading were not considered when establishing the composite rating.

<sup>d</sup>The DOK distribution of the grade 11-12 standards were used for all four assessment programs for comparison purposes; research by WestEd found the DOK distribution of the grade 9-10 standards was not substantively different from the DOK distribution of the grade 11-12 standards (WestEd. (2011). Smarter Balanced Assessment Consortium Common Core State Standards Analysis: Eligible Content for the Summative Assessment. Prepared for the Smarter Balanced Assessment Consortium: Edynn Sato, Rachel Lagunoff, and Peter Worth).

## MCAS – HIGH SCHOOL MATHEMATICS SUMMATIVE ASSESSMENT

### OVERALL SUMMARY

MCAS received a “Good Match” on **Content** for its high school mathematics summative assessment. As recommended by the criteria, at least half the score points on this assessment were aligned exclusively to prerequisites for careers and a wide range of postsecondary studies. However, reviewers noted that some domains/standards were assessed multiple times while other domains/standards were not assessed at all. Although concepts, procedures, and applications were each addressed on this assessment, the recommended balance among the three categories was not met. Additionally, of the items that assessed conceptual understanding, reviewers perceived the complexity of those items to be at a very low level. Further, items that assessed application did not require the student to use context to determine meaning or to answer the item, as recommended by the criteria.

MCAS received a “Limited Match” on **Depth** for its high school mathematics summative assessment. None of the items on this assessment specified a Mathematical Practice; to meet this criterion, items need to assess Mathematical Practices and content. The distribution of the cognitive demand of this assessment was not balanced appropriately with the distribution of the cognitive demand of the standards, as recommended by the criteria; reviewers found the distribution of cognitive demand of this assessment only partially matched the distribution of cognitive demand of the standards as a whole; reviewers found there was too much coverage of the lower levels of cognitive demand. The items on this assessment were generally free of technical and editorial issues, and they were free of bias. Per the criteria, various item types were represented and one of those types required students to generate a response.

## MCAS – HIGH SCHOOL MATHEMATICS SUMMATIVE ASSESSMENT<sup>a</sup>

### I. CONTENT: Assesses the content most needed for College and Career Readiness

G

As recommended by the criteria, at least half the score points on this assessment were aligned exclusively to prerequisites for careers and a wide range of postsecondary studies. However, reviewers noted that some domains/standards were assessed multiple times while other domains/standards were not assessed at all. Although concepts, procedures, and applications were each addressed on this assessment, the recommended balance among the three categories was not met. Additionally, of the items that assessed conceptual understanding, reviewers perceived the complexity of those items to be at a very low level. Further, items that assessed application did not require the student to use context to determine meaning or to answer the item, as recommended by the criteria.

Criteria	Rating	Group Summary Statement
<b><u>C.1 Focus.</u></b> <sup>b</sup> Tests focus strongly on content most needed in each grade or course for success in later mathematics (prerequisites for careers and a wide range of postsecondary studies).	<span style="background-color: #90EE90; border-radius: 50%; padding: 2px 6px;">G</span>	As recommended by the criteria, at least half the score points on this assessment were aligned exclusively to prerequisites for careers and a wide range of postsecondary studies. However, reviewers noted that some domains/standards were assessed multiple times while other domains/standards were not assessed at all.
C.2 Concepts, procedures, and applications. Place balanced emphasis on measurement of conceptual understanding, fluency and procedural skill, and application of mathematics.	<span style="background-color: #FFD700; border-radius: 50%; padding: 2px 6px;">L</span>	Although concepts, procedures, and applications were each addressed on this assessment, the recommended balance among the three categories was not met. Additionally, of the items that assessed conceptual understanding, reviewers perceived the complexity of those items to be at a very low level. Further, items that assessed application did not require the student to use context to determine meaning or to answer the item, as recommended by the criteria.

<sup>a</sup> Legend: E = Excellent, G = Good, L = Limited, W = Weak, IE = Insufficient Evidence



<sup>b</sup> Underlined criteria indicate criteria that are to have more weight in the composite rating.

## MCAS – HIGH SCHOOL MATHEMATICS SUMMATIVE ASSESSMENT<sup>a</sup>

### II. DEPTH: Assesses the depth that reflect the demands of College and Career Readiness



None of the items on this assessment specified a Mathematical Practice; to meet this criterion, items need to assess Mathematical Practices and content, so there was insufficient evidence (IE) to provide a rating. The distribution of the cognitive demand of this assessment was not balanced appropriately with the distribution of the cognitive demand of the standards, as recommended by the criteria; reviewers found the distribution of cognitive demand of this assessment only partially matched the distribution of cognitive demand of the standards as a whole. Specifically, reviewers found there was too much coverage of the lower levels of cognitive demand. The items on this assessment were generally free of technical and editorial issues, and they were free of bias. Per the criteria, various item types were represented and one of those types required students to generate a response.

Criteria	Rating	Group Summary Statement
<b><u>C.3 Connecting practice to content.</u></b> <sup>b,c</sup> Questions meaningfully connect mathematical practices and processes with mathematical content.	IE	None of the items on this assessment specified a Mathematical Practice; to meet this criterion, items need to assess Mathematical Practices and content, so there was insufficient evidence (IE) to provide a rating. Thus reviewers did not consider Criterion C.3 when establishing the composite Depth rating (indicated by the gray shading).
C.4 Cognitive demand. Distribution of cognitive demand for each grade level is sufficient to assess the depth and complexity of the standards.		The distribution of the cognitive demand of this assessment was not balanced appropriately with the distribution of the cognitive demand of the standards, as recommended by the criteria; reviewers found the distribution of cognitive demand of this assessment only partially matched the distribution of cognitive demand of the standards as a whole. Specifically, reviewers found there was too much coverage of the lower levels of cognitive demand.  It should be noted that assessment programs implement different cognitive complexity frameworks that might impact this rating, especially those programs that do not use Webb's DOK methodology to determine cognitive demand, as used in this study. MCAS uses the National Assessment of Educational Progress (NAEP) model for cognitive complexity.
C.5 High-quality items and a variety of item types. Items are of high technical and editorial quality and each test form includes at least two item types including at least one that requires students to generate a response.		The items on this assessment were generally free of technical and editorial issues, and they were free of bias. Per the criteria, various item types were represented and one of those types required students to generate a response.

<sup>a</sup> Legend: E = Excellent, G = Good, L = Limited, W = Weak, IE = Insufficient Evidence

<sup>b</sup> Underlined criteria indicate criteria that are to have more weight in the composite rating.

<sup>c</sup> Cells that have gray shading were not considered when establishing the composite rating.

## Appendix I: PARCC Criteria B and C Ratings and Summary Statements

### PARCC – HIGH SCHOOL ENGLISH LANGUAGE ARTS/LITERACY SUMMATIVE ASSESSMENT

#### OVERALL SUMMARY

PARCC received an “Excellent match” on **Content** for its high school ELA/literacy summative assessment. Nearly all of the items on this assessment required close reading and analysis of the text. The items also focused on central ideas/themes and important particulars. Most items were text dependent and they were aligned to the specifics of the standard. As recommended by the criteria, students were required to provide textual evidence in their responses to most items. As also recommended by the criteria, all three writing types (expository, persuasive/argumentative, and narrative) were represented on this assessment. All writing prompts required writing to relevant sources. Additionally, students were required to support, infer, and draw conclusions to support their claims. As recommended by the criteria, the large majority of vocabulary items on this assessment focused on tier 2 words (that is, words commonly used in written texts, often referred to as “general academic words”) and required students to use context to determine meaning. Additionally, the large majority of items that measured language skills emphasized the conventions most important for readiness and mirrored real world skills and tasks. Per the criteria, vocabulary and language skills were reported as sub-scores. Reviewers judged the large majority of research items and writing prompts to require analysis, synthesis, and/or organization of information; these items also required citation of evidence.






PARCC received a “Limited match” on **Depth** for its high school ELA/literacy summative assessment. The texts in this assessment were of high quality and used open sources, but reviewers judged them to be overly rigorous. A larger range of text structure and purposes were needed to meet the criteria; specifically, reviewers found that less than half of the passages on this assessment were informational, while the criteria recommended about two-thirds of the texts be informational. Of the passages that were informational, the large majority was expository; however, the informational texts did not represent literary nonfiction, history/social science, and science/technical and, thus, the three categories were not evenly split, as recommended by the criteria. Per the criteria, quantitative and qualitative measures should be used to place each text at the appropriate grade band and level. Reviewers found the distribution of cognitive demand on this assessment only partially matched the distribution of cognitive demand of the standards. Reviewers felt there were a lot of items at high DOK levels and not enough items at the lower levels, making it difficult to adequately assess the full range of student abilities. Per the criteria, at least two item types were included on this assessment and one of those item types required students to generate a response. Reviewers found the technology enhanced items were varied, and they required a variety of student skills and tasks. As recommended by the criteria, all or nearly all items reflected technical quality, editorial accuracy, and alignment to standards.

## PARCC – HIGH SCHOOL ENGLISH LANGUAGE ARTS/LITERACY SUMMATIVE ASSESSMENT<sup>a</sup>

### I. CONTENT: Assesses the content most needed for College and Career Readiness

E

Nearly all of the items on this assessment required close reading and analysis of the text. The items also focused on central ideas/themes and important particulars. Most items were text dependent and they were aligned to the specifics of the standard. As recommended by the criteria, students were required to provide textual evidence in their responses to most items. As recommended by the criteria, all three writing types (expository, persuasive/argumentative, and narrative) were represented on this assessment. All writing prompts required writing to relevant sources. Students were required to support, infer, and draw conclusions to support their claims. As recommended by the criteria, the large majority of vocabulary items on this assessment focused on tier 2 words (that is, words commonly used in written texts, often referred to as “general academic words”) and required students to use context to determine meaning. Additionally, the large majority of items that measured language skills emphasized the conventions most important for readiness and mirrored real world skills and tasks. Per the criteria, vocabulary and language skills were reported as sub-scores. As also recommended by the criteria, reviewers judged the large majority of research items and writing prompts to require analysis, synthesis, and/or organization of information; these items also required citation of evidence.

Criteria	Rating	Group Summary Statement
<b><u>B.3 Reading</u></b> <sup>b</sup> Require students to read closely and use specific evidence from texts to obtain and defend correct responses.		Nearly all of the items on this assessment required close reading and analysis of the text. The items also focused on central ideas/themes and important particulars. Most items were text dependent and they were aligned to the specifics of the standard. As recommended by the criteria, students were required to provide textual evidence in their responses to most items.
<b><u>B.5 Writing</u></b> . Require students to engage in close reading and analysis of texts. Across grade band, tests include balance of expository, persuasive/argument, and narrative writing.		As recommended by the criteria, all three writing types (expository, persuasive/argumentative, and narrative) were represented on this assessment. All writing prompts required writing to relevant sources. Additionally, students were required to support, infer, and draw conclusions to support their claims.  Note: All items that were aligned to a writing standard were included in the evaluation of Criterion B.5, regardless of whether the item required students to actually generate a written response.
B.6 Vocabulary and language skills. Place sufficient emphasis on academic vocabulary and language conventions used in real-world activities.		As recommended by the criteria, the large majority of vocabulary items on this assessment focused on tier 2 words (that is, words commonly used in written texts, often referred to as “general academic words”) and required students to use context to determine meaning. Additionally, the large majority of items that measured language skills emphasized the conventions most important for readiness and mirrored real world skills and tasks. Per the criteria, vocabulary and language skills were reported as sub-scores.
B.7 Research and inquiry. Require students to demonstrate the ability to find, process, synthesize and organize information from multiple sources.		As recommended by the criteria, reviewers judged the large majority of research items and writing prompts to require analysis, synthesis, and/or organization of information; these items also required citation of evidence.
B.8 Speaking and listening. <sup>c</sup> Over time and as advances allow, measure speaking and listening skills.		None of the items assessed speaking and listening skills required for college and career readiness; the criteria recommend assessing Speaking and Listening skills over time and as advances allow. Thus, this criterion was not included when establishing the composite Content rating (indicated by gray shading).

<sup>a</sup> Legend: E = Excellent, G = Good, L = Limited, W = Weak, IE = Insufficient Evidence

<sup>b</sup> Underlined criteria indicate criteria that are to have more weight in the composite rating.

<sup>c</sup> Cells that have gray shading were not considered when establishing the composite rating.



## PARCC – HIGH SCHOOL ENGLISH LANGUAGE ARTS/LITERACY SUMMATIVE ASSESSMENT<sup>a</sup>



L

### II. DEPTH: Assesses the depth that reflect the demands of College and Career Readiness

The texts in this assessment were of high quality and used open sources, but reviewers judged them to be overly rigorous. A larger range of text structure and purposes were needed to meet the criteria; specifically, reviewers found that less than half of the passages on this assessment were informational, while the criteria recommend about two-thirds of the texts be informational. Of the passages that were informational, the large majority was expository; however, the informational texts did not represent literary nonfiction, history/social science, and science/technical and, thus, the three categories were not evenly split, as recommended by the criteria. Reviewers found the distribution of cognitive demand on this assessment only partially matched the distribution of cognitive demand of the standards. Reviewers felt there were a lot of items at high DOK levels and not enough items at the lower levels, making it difficult to adequately assess the full range of student abilities. Per the criteria, at least two item types were included on this assessment and one of those item types required students to generate a response. Reviewers found the technology enhanced items were varied, and they required a variety of student skills and tasks. As recommended by the criteria, all or nearly all items reflected technical quality, editorial accuracy, and alignment to standards.

Criteria	Rating	Group Summary Statement
<p><b><u>B.1 Text quality and types.</u></b><sup>b</sup> Include aligned balance of high-quality literary and informational texts.</p>	<div style="border: 1px solid black; border-radius: 50%; width: 30px; height: 30px; display: flex; align-items: center; justify-content: center; background-color: #ffc107;">L</div>	<p>The texts in this assessment were of high quality and used open sources, but reviewers judged them to be overly rigorous. A larger range of text structure and purposes were needed to meet the criteria; specifically, reviewers found that less than half of the passages on this assessment were informational, while the criteria recommend about two-thirds of the texts be informational. Of the passages that were informational, the large majority was expository; however, the informational texts did not represent literary nonfiction, history/social science, and science/technical and, thus, the three categories were not evenly split, as recommended by the criteria.</p> <p>It should be noted there are typically only a limited number of passages that can be included on any given assessment, thus, the methodology’s recommended distribution of passage types could be influenced greatly by a single discrepancy that might result in a different (lower or higher) rating.</p>
<p><b><u>B.2 Complexity of texts.</u></b><sup>c</sup> Passages are at appropriate levels of text complexity, increasing through the grades, and multiple forms of authentic, high-quality texts are used.</p>	<div style="border: 1px solid black; border-radius: 50%; width: 30px; height: 30px; display: flex; align-items: center; justify-content: center; background-color: #6c757d;">G</div>	<p>Per the Criteria, quantitative and qualitative measures should be used to place each text at the appropriate grade band and level. The PARCC program documentation indicated the use of both quantitative and qualitative measures of text complexity; however, reviewers could not provide a rating based on the items because it was not possible to obtain complexity metadata from all programs included in this study in a format for the reviewers to evaluate.</p> <p>Note: The Criterion B.2 rating is based solely on program documentation as reviewers were not able to rate the extent to which quantitative measures are used to place each text in a grade band. Thus, reviewers did not consider the Criterion B.2 rating when establishing the composite Depth rating (indicated by the gray shading).</p>

## PARCC – HIGH SCHOOL ENGLISH LANGUAGE ARTS/LITERACY SUMMATIVE ASSESSMENT

Criteria	Rating	Group Summary Statement
<p>B.4 Cognitive demand.<sup>d</sup> Distribution of cognitive demand for each grade level is sufficient to assess the depth and complexity of the standards.</p>		<p>Reviewers found the distribution of cognitive demand on this assessment only partially matched the distribution of cognitive demand of the standards. Reviewers felt there were a lot of items at high DOK levels and not enough items at the lower levels, making it difficult to adequately assess the full range of student abilities.</p> <p>It should be noted that assessment programs implement different cognitive complexity frameworks that might impact this rating, especially those programs that do not use Webb’s DOK methodology to determine cognitive demand, as used in this study. PARCC uses the Cognitive Complexity Framework.<sup>e</sup></p>
<p>B.9 High-quality items and a variety of item types. Items are of high technical and editorial quality and each test form includes at least two item types including at least one that requires students to generate a response.</p>		<p>Per the criteria, at least two item types were included on this assessment and one of those item types required students to generate a response. Reviewers found the technology enhanced items were varied, and they required a variety of student skills and tasks. As recommended by the criteria, all or nearly all items reflected technical quality, editorial accuracy, and alignment to standards.</p>

<sup>a</sup> Legend: E = Excellent, G = Good, L = Limited, W = Weak, IE = Insufficient Evidence

<sup>b</sup> Underlined criteria indicate criteria that are to have more weight in the composite rating.

<sup>c</sup> Cells that have gray shading were not considered when establishing the composite rating.

<sup>d</sup>The DOK distribution of the grade 11-12 standards were used for all four assessment programs for comparison purposes; research by WestEd found the DOK distribution of the grade 9-10 standards was not substantively different from the DOK distribution of the grade 11-12 standards (WestEd. (2011). Smarter Balanced Assessment Consortium Common Core State Standards Analysis: Eligible Content for the Summative Assessment. Prepared for the Smarter Balanced Assessment Consortium: Edynn Sato, Rachel Lagunoff, and Peter Worth).

<sup>e</sup>PARCC developed the Cognitive Complexity Framework, which recognizes that text complexity and item/task complexity interact to determine the overall complexity of a task.

## PARCC – HIGH SCHOOL MATHEMATICS SUMMATIVE ASSESSMENT<sup>a</sup>

### OVERALL SUMMARY

PARCC received an “Excellent Match” for **Content** on its high school mathematics summative assessment. As recommended by the criteria, at least half of the score points on this assessment were aligned to the widely applicable prerequisites for careers and a wide range of postsecondary studies. The items aligned well to high school content; reviewers found that nearly all domains/standards within the widely applicable prerequisites were assessed. Additionally, all content was at grade level and it was reflective of student success at the high school level. Although the distribution of score points that assessed conceptual understanding, procedural skills, and application was not equally balanced as recommended by the criteria, reviewers judged the items that did assess application to be rich in content and practice.

PARCC received a “Good Match” for **Depth** on its high school mathematics summative assessment. As recommended by the criteria, all items that assessed a Mathematical Practice also aligned to at least one standard. Reviewers found the distribution of the DOK of the items on this assessment was similar but did not fully match the distribution of the DOK of the standards. Reviewers believed more items were needed at the higher DOK levels. As recommended by the criteria, this assessment included a variety of item types and one of those types required students to generate rather than select a response. Reviewers judged the items to be aligned to the standards and technically accurate.

### I. CONTENT: Assesses the content most needed for College and Career Readiness

**E**

As recommended by the criteria, at least half of the score points on this assessment were aligned to the widely applicable prerequisites for careers and a wide range of postsecondary studies. The items aligned well to high school content; reviewers found that nearly all domains/standards within the widely applicable prerequisites were assessed. Additionally, all content was at grade level and it was reflective of student success at the high school level. Although the distribution of score points that assessed conceptual understanding, procedural skills, and application was not equally balanced as recommended by the criteria, reviewers judged the items that did assess application to be rich in content and practice.

Criteria	Rating	Group Summary Statement
<b><u>C.1 Focus</u></b> <sup>b</sup> Tests focus strongly on content most needed in each grade or course for success in later mathematics (prerequisites for careers and a wide range of postsecondary studies).	<b>E</b>	As recommended by the criteria, at least half of the score points on this assessment were aligned to the widely applicable prerequisites for careers and a wide range of postsecondary studies. The items aligned well to high school content; reviewers found that nearly all domains/standards within the widely applicable prerequisites were assessed. Additionally, all content was at grade level and it was reflective of student success at the high school level.
C.2 Concepts, procedures, and applications. Place balanced emphasis on measurement of conceptual understanding, fluency and procedural skill, and application of mathematics.	<b>G</b>	Although the distribution of score points that assessed conceptual understanding, procedural skills, and application was not equally balanced as recommended by the criteria, reviewers judged the items that did assess application to be rich in content and practice.

<sup>a</sup> Legend: E = Excellent, G = Good, L = Limited, W = Weak, IE = Insufficient Evidence

<sup>b</sup> Underlined criteria indicate criteria that are to have more weight in the composite rating.

## PARCC – HIGH SCHOOL MATHEMATICS SUMMATIVE ASSESSMENT<sup>a</sup>

### II. DEPTH: Assesses the depth that reflect the demands of College and Career Readiness

G

As recommended by the criteria, all items that assessed a Mathematical Practice also aligned to at least one standard. Reviewers found the distribution of the DOK of the items on this assessment was similar but did not fully match the distribution of the DOK of the standards. Reviewers believed more items were needed at the higher DOK levels. As recommended by the criteria, this assessment included a variety of item types and one of those types required students to generate rather than select a response. Reviewers judged the items to be aligned to the standards and technically accurate.

Criteria	Rating	Group Summary Statement
<b>C.3 Connecting practice to content.</b> <sup>b</sup> Questions meaningfully connect mathematical practices and processes with mathematical content.	<span style="border: 1px solid green; border-radius: 50%; width: 30px; height: 30px; display: flex; align-items: center; justify-content: center; margin: 0 auto;">E</span>	As recommended by the criteria, all items that assessed a Mathematical Practice also aligned to at least one standard.
C.4 Cognitive demand. Distribution of cognitive demand for each grade level is sufficient to assess the depth and complexity of the standards.	<span style="border: 1px solid green; border-radius: 50%; width: 30px; height: 30px; display: flex; align-items: center; justify-content: center; margin: 0 auto;">G</span>	Reviewers found the distribution of the DOK of the items on this assessment was similar but did not fully match the distribution of the DOK of the standards. Reviewers believed more items were needed at the higher DOK levels.  It should be noted that assessment programs implement different cognitive complexity frameworks that might impact this rating, especially those programs that do not use Webb’s DOK methodology to determine cognitive demand, as used in this study. PARCC uses the Cognitive Complexity Framework. <sup>c</sup>
C.5 High-quality items and a variety of item types. Items are of high technical and editorial quality and each test form includes at least two item types including at least one that requires students to generate a response.	<span style="border: 1px solid green; border-radius: 50%; width: 30px; height: 30px; display: flex; align-items: center; justify-content: center; margin: 0 auto;">E</span>	As recommended by the <i>Criteria</i> , this assessment included a variety of item types and one of those types required students to generate rather than select a response. Reviewers judged the items to be aligned to the standards and technically accurate.

<sup>a</sup> Legend: E = Excellent, G = Good, L = Limited, W = Weak, IE = Insufficient Evidence

<sup>b</sup> Underlined criteria indicate criteria that are to have more weight in the composite rating.

<sup>c</sup> PARCC developed the Cognitive Complexity Framework, which recognizes that text complexity and item/task complexity interact to determine the overall complexity of a task.

## Appendix J: Smarter Balanced Criteria B and C Ratings and Summary Statements

**SMARTER BALANCED – HIGH SCHOOL ENGLISH LANGUAGE ARTS/LITERACY SUMMATIVE ASSESSMENT****OVERALL SUMMARY**

Smarter Balanced received an “Excellent Match” for **Content** on its high school ELA/literacy summative assessment. As recommended by the criteria, the large majority of items on this assessment required close reading and analysis of text. Nearly all items were aligned to the specifics of the standards. The majority of items also focused on the central ideas and important particulars rather than superficial or peripheral concepts. Most items required students to interact with the text to produce responses, as recommended by the criteria; more than half the reading score points on this assessment required direct use of textual evidence. This assessment slightly emphasized expository writing types, but it included expository and argumentative/persuasive writing prompts. Most writing items required students to generate a response; these prompts were text-based and required students to write to sources, which met requirements of the criteria. There were some writing items that did not require students to generate a response, but instead required students to evaluate and choose the correct answer from multiple responses. Although the score points devoted to assessing vocabulary was somewhat low, reviewers judged the overall assessment to strongly assess vocabulary. The large majority of items that assessed vocabulary focused on tier 2 words (that is, words commonly used in written texts, often referred to as “general academic words”) and required students to use context to determine meaning. At least three-fourths of the items on this assessment (as recommended by the criteria) mirrored real-world activities, focused on common errors, and emphasized the conventions most important for readiness. Additionally, per the criteria, language skills were reported as a sub-score. Meeting requirements of the criteria, reviewers judged all of the items that assessed research and inquiry to mirror real-world activities. Additionally, at least three-fourths of the research items required students to analyze, synthesize, and/or organize information. Items that assessed listening skills required students to take notes on main ideas and elaborate on remarks of others.

Smarter Balanced received an “Excellent Match” for **Depth** on its high school ELA/literacy summative assessment. All passages on this assessment were of high quality, and both forms included appropriately complex and interesting passages. As recommended by the criteria, this assessment emphasized informational rather than expository text. However, rather than a nearly even split among literary nonfiction, history/social science, and science/technical texts (as recommended by the criteria), reviewers found slightly less focus on history/social science texts. Reviewers found the distribution of cognitive demand on both forms matched the distribution of cognitive demand of the standards as a whole. Additionally as recommended by the criteria, reviewers found the distribution of cognitive demand of the assessment matched the higher cognitive demand (DOK3+) of the standards. As recommended by the criteria, this assessment included a variety of item types and at least one of those required students to generate rather than select a response. Additionally, reviewers found that nearly all of the items aligned well to the standards, and they reflected technical quality and editorial accuracy.




## SMARTER BALANCED – HIGH SCHOOL ENGLISH LANGUAGE ARTS/LITERACY SUMMATIVE ASSESSMENT<sup>a</sup>

### I. CONTENT: Assesses the content most needed for College and Career Readiness

E

As recommended by the criteria, the large majority of items on this assessment required close reading and analysis of text. Nearly all items were aligned to the specifics of the standards. The majority of items also focused on the central ideas and important particulars rather than superficial or peripheral concepts. Most items required students to interact with the text to produce responses, as recommended by the criteria; more than half the reading score points on this assessment required direct use of textual evidence. This assessment slightly emphasized expository writing types, but it included expository and argumentative/persuasive writing prompts. Most writing items required students to generate a response; these prompts were text-based and required students to write to sources, which met requirements of the criteria. There were some writing items that did not require students to generate a response, but instead required students to evaluate and choose the correct answer from multiple responses. Although the score points devoted to assessing vocabulary was somewhat low, reviewers judged the overall assessment to strongly assess vocabulary. The large majority of items that assessed vocabulary focused on tier 2 words (that is, words commonly used in written texts, often referred to as "general academic words") and required students to use context to determine meaning. At least three-fourths of the items on this assessment (as recommended by the criteria) mirrored real-world activities, focused on common errors, and emphasized the conventions most important for readiness. Additionally, per the criteria, language skills were reported as a sub-score. Meeting requirements of the criteria, reviewers judged all of the items that assessed research and inquiry to mirror real-world activities. Additionally, at least three-fourths of the research items required students to analyze, synthesize, and/or organize information. Items that assessed listening skills required students to take notes on main ideas and elaborate on remarks of others.

Criteria	Rating	Group Summary Statement
<b>B.3 Reading.</b> <sup>b</sup> Require students to read closely and use specific evidence from texts to obtain and defend correct responses.	<span style="border: 1px solid green; border-radius: 50%; width: 30px; height: 30px; display: flex; align-items: center; justify-content: center;">E</span>	As recommended by the criteria, the large majority of items on this assessment required close reading and analysis of text. Nearly all items were aligned to the specifics of the standards. The majority of items also focused on the central ideas and important particulars rather than superficial or peripheral concepts. Most items required students to interact with the text to produce responses, as recommended by the criteria; more than half of reading score points required direct use of textual evidence.
<b>B.5 Writing.</b> Require students to engage in close reading and analysis of texts. Across grade band, tests include balance of expository, persuasive/argument, and narrative writing.	<span style="border: 1px solid green; border-radius: 50%; width: 30px; height: 30px; display: flex; align-items: center; justify-content: center;">E</span>	<p>This assessment slightly emphasized expository writing types, but it included expository and argumentative/persuasive writing prompts. Most writing items required students to generate a response; these prompts were text-based and required students to write to sources, which met requirements of the criteria. There were some writing items that did not require students to generate a response, but instead required students to evaluate and choose the correct answer from multiple responses.</p> <p>Note: All items that were aligned to a writing standard were included in the evaluation of Criterion B.5, regardless of whether the item required students to actually generate a written response.</p>

SMARTER BALANCED – HIGH SCHOOL ENGLISH LANGUAGE ARTS/LITERACY SUMMATIVE ASSESSMENT		
Criteria	Rating	Group Summary Statement
B.6 Vocabulary and language skills. Place sufficient emphasis on academic vocabulary and language conventions used in real-world activities.		Although the score points devoted to assessing vocabulary was somewhat low, reviewers judged the overall assessment to strongly assess vocabulary. The large majority of items that assessed vocabulary focused on tier 2 words (that is, words commonly used in written texts, often referred to as "general academic words") and required students to use context to determine meaning. At least three-fourths of the items on this assessment (as recommended by the criteria) mirrored real-world activities, focused on common errors, and emphasized the conventions most important for readiness. Additionally, per the criteria, language skills were reported as a sub-score.
B.7 Research and inquiry. Require students to demonstrate the ability to find, process, synthesize and organize information from multiple sources.		Meeting requirements of the criteria, reviewers judged all of the items that assessed research and inquiry to mirror real-world activities. Additionally, at least three-fourths of the research items required students to analyze, synthesize, and/or organize information.
B.8 Speaking and listening. <sup>c</sup> Over time and as advances allow, measure speaking and listening skills.		Items that assessed listening skills were based on texts and other stimuli that met the criteria for complexity, range, and quality. These items also permitted the evaluation of active listening skills such as taking notes on main ideas and elaborating on remarks of others. Speaking skills were not assessed on this assessment. The criteria recommend assessing Speaking and Listening skills over time and as advances allow. Thus, this criterion was not included when establishing the composite Content rating (indicated by gray shading).

<sup>a</sup> Legend: E = Excellent, G = Good, L = Limited, W = Weak, IE = Insufficient Evidence

<sup>b</sup> Underlined criteria indicate criteria that are to have more weight in the composite rating.

<sup>c</sup> Cells that have gray shading were not considered when establishing the composite rating.

## SMARTER BALANCED – HIGH SCHOOL ENGLISH LANGUAGE ARTS/LITERACY SUMMATIVE ASSESSMENT<sup>a</sup>

### II. DEPTH: Assesses the depth that reflect the demands of College and Career Readiness



E

All passages on this assessment were of high quality, and the assessment included appropriately complex and interesting passages. As recommended by the criteria, there was an emphasis on informational rather than expository text; however, rather than a nearly even split among literary nonfiction, history/social science, and science/technical texts (as recommended by the criteria), reviewers found slightly less focus on history/social science texts. Reviewers found the distribution of cognitive demand on both forms matched the distribution of cognitive demand of the standards as a whole. Additionally as recommended by the criteria, reviewers found the distribution of cognitive demand of the assessment matched the higher cognitive demand (DOK3+) of the standards. As recommended by the criteria, this assessment included a variety of item types and at least one of those required students to generate rather than select a response. Additionally, reviewers found that nearly all of the items aligned well to the standards, and they reflected technical quality and editorial accuracy.

Criteria	Rating	Group Summary Statement
<p><b><u>B.1 Text quality and types.</u></b><sup>b</sup> Include aligned balance of high-quality literary and informational texts.</p>	<div style="border: 1px solid green; border-radius: 50%; width: 30px; height: 30px; display: flex; align-items: center; justify-content: center; margin: 0 auto;">E</div>	<p>All passages on this assessment were of high quality, and the assessment included appropriately complex and interesting passages. As recommended by the criteria, there was an emphasis on informational rather than expository text; however, rather than a nearly even split among literary nonfiction, history/social science, and science/technical texts (as recommended by the criteria), reviewers found slightly less focus on history/social science texts.</p> <p>It should be noted there are typically only a limited number of passages that can be included on any given assessment, thus, the methodology’s recommended distribution of passage types could be influenced greatly by a single discrepancy that might result in a different (lower or higher) rating.</p>
<p><b><u>B.2 Complexity of texts.</u></b><sup>c</sup> Passages are at appropriate levels of text complexity, increasing through the grades, and multiple forms of authentic, high-quality texts are used.</p>	<div style="border: 1px solid gray; border-radius: 50%; width: 30px; height: 30px; display: flex; align-items: center; justify-content: center; margin: 0 auto;">G</div>	<p>Per the criteria, quantitative and qualitative measures should be used to place each text at the appropriate grade band and level. The Smarter Balanced program documentation indicated the use of both quantitative and qualitative measures of text complexity for stimuli selected; however, reviewers could not provide a rating based on the items because it was not possible to obtain complexity metadata from all programs included in this study in a format for the reviewers to evaluate.</p> <p>Note: The Criterion B.2 rating is based solely on program documentation as reviewers were not able to rate the extent to which quantitative measures are used to place each text in a grade band. Thus, reviewers did not consider the Criterion B.2 rating when establishing the composite Depth rating (indicated by the gray shading).</p>



## SMARTER BALANCED – HIGH SCHOOL ENGLISH LANGUAGE ARTS/LITERACY SUMMATIVE ASSESSMENT

Criteria	Rating	Group Summary Statement
B.4 Cognitive demand. <sup>d</sup> Distribution of cognitive demand for each grade level is sufficient to assess the depth and complexity of the standards.		<p>Reviewers found the distribution of cognitive demand on this assessment matched the distribution of cognitive demand of the standards as a whole. Additionally as recommended by the criteria, reviewers found the distribution of cognitive demand of the assessment matched the higher cognitive demand (DOK3+) of the standards.</p> <p>It should be noted that assessment programs implement different cognitive complexity frameworks that might impact this rating, especially those programs that do not use Webb's DOK methodology to determine cognitive demand, as used in this study. Smarter Balanced uses Karen Hess' Cognitive Rigor Matrix, which involves Webb's DOK methodology.</p>
B.9 High-quality items and a variety of item types. Items are of high technical and editorial quality and each test form includes at least two item types including at least one that requires students to generate a response.		<p>As recommended by the criteria, this assessment included a variety of item types and at least one of those required students to generate rather than select a response. Additionally, reviewers found that nearly all of the items aligned well to the standards, and they reflected technical quality and editorial accuracy.</p>

<sup>a</sup> Legend: E = Excellent, G = Good, L = Limited, W = Weak, IE = Insufficient Evidence

<sup>b</sup> Underlined criteria indicate criteria that are to have more weight in the composite rating.

<sup>c</sup> Cells that have gray shading were not considered when establishing the composite rating.

<sup>d</sup>The DOK distribution of the grade 11-12 standards were used for all four assessment programs for comparison purposes; research by WestEd found the DOK distribution of the grade 9-10 standards was not substantively different from the DOK distribution of the grade 11-12 standards (WestEd. (2011). Smarter Balanced Assessment Consortium Common Core State Standards Analysis: Eligible Content for the Summative Assessment. Prepared for the Smarter Balanced Assessment Consortium: Edynn Sato, Rachel Lagunoff, and Peter Worth).

## SMARTER BALANCED – HIGH SCHOOL MATHEMATICS SUMMATIVE ASSESSMENT<sup>a</sup>

### OVERALL SUMMARY

Smarter Balanced received an “Excellent Match” for **Content** on its high school mathematics summative assessment. As recommended by the criteria, at least half the score points on this assessment aligned exclusively to prerequisites for careers and a wide variety of postsecondary studies. Additionally, most of the domains/standards within the widely applicable prerequisites were assessed. This assessment included items that assessed conceptual understanding, procedural skill and fluency, and application; however, only one form that was reviewed had a balance of the three, as recommended by the criteria.

Smarter Balanced received an “Excellent Match” for **Depth** on its high school mathematics summative assessment. Reviewers believed the items on this assessment meaningfully connected practice to content. Per the criteria, all of the items that assessed a Mathematical Practice also aligned to at least one content standard. Reviewers found the distribution of cognitive demand of the assessment matched the distribution of cognitive demand of the standards as a whole. As recommended by the criteria, reviewers found the percentage of score points on this assessment matched the higher cognitive demand (DOK 3+) of the standards. As recommended by the criteria, this assessment included at least two item types and one of them required students to generate rather than select a response. Reviewers found most items aligned well to the standards, and they reflected technical quality and editorial accuracy; however, reviewers judged a number of the constructed response items to have excessive verbiage and certain contexts required prior knowledge.

### I. CONTENT: Assesses the content most needed for College and Career Readiness E

As recommended by the criteria, at least half the score points on this assessment aligned exclusively to prerequisites for careers and a wide variety of postsecondary studies. Additionally, most of the domains/standards within the widely applicable prerequisites were assessed. This assessment included items that assessed conceptual understanding, procedural skill and fluency, and application; however, only one form that was reviewed had a balance of the three, as recommended by the criteria.

Criteria	Rating	Group Summary Statement
<b><u>C.1 Focus.</u></b> <sup>b</sup> Tests focus strongly on content most needed in each grade or course for success in later mathematics (prerequisites for careers and a wide range of postsecondary studies).	<span style="border: 1px solid green; border-radius: 50%; padding: 5px; font-weight: bold; font-size: 1.5em; color: white;">E</span>	As recommended by the criteria, at least half the score points on this assessment aligned exclusively to prerequisites for careers and a wide variety of postsecondary studies. Additionally, most of the domains/standards within the widely applicable prerequisites were assessed.
C.2 Concepts, procedures, and applications. Place balanced emphasis on measurement of conceptual understanding, fluency and procedural skill, and application of mathematics.	<span style="border: 1px solid green; border-radius: 50%; padding: 5px; font-weight: bold; font-size: 1.5em; color: white;">G</span>	This assessment included items that assessed conceptual understanding, procedural skill and fluency, and application; however, only one form that was reviewed had a balance of the three, as recommended by the criteria.

<sup>a</sup> Legend: E = Excellent, G = Good, L = Limited, W = Weak, IE = Insufficient Evidence




<sup>b</sup> Underlined criteria indicate criteria that are to have more weight in the composite rating.

## SMARTER BALANCED – HIGH SCHOOL MATHEMATICS SUMMATIVE ASSESSMENT<sup>a</sup>

E

### II. DEPTH: Assesses the depth that reflect the demands of College and Career Readiness

Reviewers believed the items on this assessment meaningfully connected practice to content. Per the criteria, all of the items that assessed a Mathematical Practice also aligned to at least one content standard. Reviewers found the distribution of cognitive demand of this assessment matched the distribution of cognitive demand of the standards as a whole. As recommended by the criteria, reviewers found the percentage of score points for this assessment matched the higher cognitive demand (DOK 3+) of the standards. As recommended by the criteria, this assessment included at least two item types and one of them required students to generate rather than select a response. Reviewers found most items aligned well to the standards, and they reflected technical quality and editorial accuracy; however, reviewers judged a number of the constructed response items to have excessive verbiage and certain contexts required prior knowledge.

Criteria	Rating	Group Summary Statement
<b><u>C.3 Connecting practice to content.</u></b> <sup>b</sup> Questions meaningfully connect mathematical practices and processes with mathematical content.		Reviewers believed the items on this assessment meaningfully connected practice to content. Per the criteria, all of the items that assessed a Mathematical Practice also aligned to at least one content standard.
C.4 Cognitive demand. Distribution of cognitive demand for each grade level is sufficient to assess the depth and complexity of the standards.		Reviewers found the distribution of cognitive demand of this assessment matched the distribution of cognitive demand of the standards as a whole. As recommended by the criteria, reviewers found the percentage of score points for this assessment matched the higher cognitive demand (DOK 3+) of the standards.  It should be noted that assessment programs implement different cognitive complexity frameworks that might impact this rating, especially those programs that do not use Webb's DOK methodology to determine cognitive demand, as used in this study. Smarter Balanced uses Karen Hess' Cognitive Rigor Matrix, which involves Webb's DOK methodology.
C.5 High-quality items and a variety of item types. Items are of high technical and editorial quality and each test form includes at least two item types including at least one that requires students to generate a response.		As recommended by the criteria, this assessment included at least two item types and one of them required students to generate rather than select a response. Reviewers found most items aligned well to the standards, and they reflected technical quality and editorial accuracy; however, reviewers judged a number of the constructed response items to have excessive verbiage and certain contexts required prior knowledge.

<sup>a</sup> Legend: E = Excellent, G = Good, L = Limited, W = Weak, IE = Insufficient Evidence

<sup>b</sup> Underlined criteria indicate criteria that are to have more weight in the composite rating.

## Appendix K: Testing Program Responses to HQAP High School Study

### *ACT Aspire Response to HQAP High School Study*

ACT used feedback from this study, along with data from operational administrations, and input from educators and ACT customers, to add to the roadmap of improvements for ACT Aspire. ACT will address the following changes:

#### **For English Language Arts**

- **ACT Aspire writing** tasks were designed to assess student writing competencies without the heavier reading load of “writing to sources” tasks. We are currently exploring designs for supplementary tasks that measure these valuable literacy skills. These tasks would further improve the match with the “Assessing research and inquiry” criterion, for which ACT Aspire received a Good rating in this study.
- As verified by this study, the **ACT Aspire reading** test requires students to refer to textual evidence. The study recommended an increase in the number of items asking students to explicitly cite that evidence. Although the test currently has constructed-response items that require students to cite evidence, these do not constitute the majority of score points. ACT is adding new technology-enhanced questions that require students to select evidence directly from reading passages in order to support claims and interpretations. Some of these are operational in 2015–2016, but ACT also continues to explore new ways to assess student use of evidence from texts.
- The study found a Limited Match for study criterion B.9 (high-quality items and variety of item types). The study report states, “Reviewers also felt the items had readability issues because students were not provided specific instructions for responding to the various item types.” Based on ACT’s observations of the study review session, these “readability” issues refer only to the **ACT Aspire English** test. Reviewers expressed specific concerns about the format of test instructions provided to the students. However, the format of the questions is based on a design used successfully for many years, at the same grade level as a part of the ACT PLAN program; this same format has been used successfully as a part of the ACT, with nearly identical instructions. ACT has found no evidence to suggest that students encounter issues with the test instructions, but ACT will review the English test instructions to ensure that they are clear.
- We emphasize that these “readability” issues were only raised about one aspect of the ELA tests, and the **ACT Aspire ELA** tests received Good Match scores for B.1 (text quality and types) and B.2 (complexity of texts), as well as the Good Match for overall Depth. Furthermore, ACT Aspire English uses a variety of item types, and reviewers found that this variety met the requirements of study criterion B.9.

#### **For Mathematics**

- Reviewers recommended ACT Aspire pursue more variety in technology-enhanced questions, and we agree. ACT research is active, recognizing that these questions must have proven advantages to justify the additional expense to schools. It is not about technology for technology’s sake.
- ACT collects independent information about what is important for student success. We will be working to increase the number of items focused on the “major work of the grade” and also gathering data to understand the balance in terms of promoting college and career readiness.

- Although the panelists believed that study criterion C.4 should be revised, ACT Aspire received a Limited Match to this criterion. The study found that too many points were at DOK 3, and ACT has a plan to decrease this level somewhat.
- We also express our agreement with the panel that the methodology's treatment of the mathematical practices needs attention. This will be a difficult task. ACT Aspire attends to the mathematical practices in a host of ways, with separate reporting categories for Modeling and for Justification and Explanation, which got no comment. In order for the project to live up to its goal of providing useful feedback, attending to the mathematical practices is essential. We also recommend more attention at the level of clusters rather than just standards, for clusters are the level of coherence for Common Core mathematics.

### For Accessibility

- Although the study panelists were unable to find information about how the program would manage providing multiple accommodations for a single student, ACT Aspire does indeed incorporate several elements of guidance about the bundling and combining of various accessibility and accommodation supports within the Aspire Accessibility User Guide. Specifically, within the chapter titled, "Administration Procedures for Accessibility Supports", we provide substantial detail regarding such considerations and describe guidance regarding multiple support combinations as it related to each permitted support. In an earlier version of the Guide, we also provided a separate section entitled, "Bundling of Supports" which addressed the issue of combining supports more broadly, but later removed this section based upon feedback that it was redundant and not needed with the detailed information provided for each support
- We accept the feedback that advice guidance regarding best practices in combining multiple accommodations is needed by the field. To this end we will be enhancing this information within the next ACT Aspire Accessibility User Guide.

### Specification Tables Showing Adjustments to ACT Aspire for 2015–2016

We have already made adjustments in the following four categories:

1. **Timing Adjustments** – Based on customer feedback and in order to allow all students a better opportunity to show what they know and can do, we will be adjusting the time per test by 5-10 minutes for all tests except (Writing will not change). (see **Table 1.1**)
2. **Adjustments to English Test** – Multiple choice items will increase for English grades 3, 4 and 5 (see **Tables 2.1 and 2.2**)
3. **Adjustments to Math Test** – Multiple choice items will increase for Mathematics grades 3, 4 and 5 and Constructed Response (CR) will decrease by a single item. (see **Tables 3.1 and 3.2**)
4. **Adjustments to Accessibility Systems** – A number of changes will be effected: Spanish Translation Forms will be expanded; color contrast palettes and the highlighter tool will be activated for online testing; custom masking, online interactive screen-reader compatibility, and an online embedded American Sign Language (ASL) video will be added; and accessibility profiles will be mapped across item interaction formats and task models for 7 learner populations. (see **Table 4.1**)

**Table 1.1.** Timing Adjustments

Grade	ACT Aspire Summative Testing Time Adjustments (in minutes)							
	English (Current)	English (New)	Math (Current)	Math (New)	Reading (Current)	Reading (New)	Science (Current)	Science (New)
3	30	40	55	65	60	65	55	60
4	30	40	55	65	60	65	55	60
5	30	40	55	65	60	65	55	60
6	35	40	60	70	60	65	55	60
7	35	40	60	70	60	65	55	60
8	35	40	65	75	60	65	55	60
*EHS	40	45	65	75	60	65	55	60

\*Early High School

**Table 2.1.** English: Number of Items by Item Type

	GRADE									
	NEW	OLD	NEW	OLD	NEW	OLD				
# of Items	3	3	4	4	5	5	6	7	8	EHS
MC	31	25	31	25	31	25	34	34	38	38
	27-28	21-22	27-28	21-22	27-28	21-22	23-25	23-25	27-29	27-29
TE	3-4	3-4	3-4	3-4	3-4	3-4	5-7	5-7	5-7	5-7

**Table 2.2.** English: Number of Points by Reporting Category

	GRADE									
	NEW	OLD	NEW	OLD	NEW	OLD				
# of Points	3	3	4	4	5	5	6	7	8	EHS
<b>Total</b>	31	25	31	25	31	25	35	35	35	50
<b>Production of Writing</b>	12-14	9-11	8-10	6-8	8-10	6-8	11-13	9-11	9-11	12-14
<b>Knowledge of Language</b>			3-5	2-4	3-5	2-4	2-4	4-6	4-6	6-8
<b>Conventions of Standard English</b>	17-19	14-16	17-19	14-16	17-19	14-16	19-21	19-21	19-21	29-31

**Table 3.1.** Mathematics: Number of Items by Item Type

		GRADE									
		NEW	OLD	NEW	OLD	NEW	OLD				
		3	3	4	4	5	5	6	7	8	EHS
MC	# of Items	30	25	30	25	30	25	34	34	38	38
		21-22	15-16	22-24	15-16	21-22	15-16	23-25	23-25	27-29	27-29
TE		5-6	5-6	5-6	5-6	5-6	5-6	5-7	5-7	5-7	5-7
CR		3	4	3	4	3	4	4	4	5	5

**Table 3.2.** Mathematics: Number of Points by Reporting Category

		GRADE									
		NEW	OLD	NEW	OLD	NEW	OLD				
		3	3	4	4	5	5	6	7	8	EHS
	# of Points										
	<b>Total</b>	39	37	39	37	39	37	46	46	53	53
	<b>Number &amp; Operations in Base 10</b>		5-7	5-8	3-5	5-8	3-5	1-3	1-3	1-3	0-2
	<b>Number &amp; Operations - Fractions</b>	3-5	2-4	6-8	4-6	6-8	4-6	1-3	1-3	1-3	0-2
	<b>The Number System</b>							3-5	3-5	2-4	1-3
	<b>Number &amp; Quantity</b>										1-3
	<b>Operations &amp; Algebraic Thinking</b>	6-8	3-5	4-6	3-5		3-5	1-3	1-3	0-2	0-2
	<b>Expressions &amp; Equations</b>							3-5	3-5	5-7	2-4
	<b>Ratios &amp; Proportional Reasoning</b>							3-5	3-5	0-2	1-3
	<b>Algebra</b>										2-4
	<b>Functions</b>									3-5	3-5
	<b>Measurement &amp; Data (measurement)</b>							0-2	0-2	1-3	1-3
	<b>Geometry</b>		3-5		3-5	4-6	3-5	5-7	4-6	6-8	5-7
	<b>Measurement &amp; Data</b>	5-7	3-5		3-5		3-5				
	<b>Measurement &amp; Data (data)</b>							0-2	1-3	1-3	1-3
	<b>Statistics &amp; Probability</b>							3-5	3-5	4-6	4-7
	<b>Justification &amp; Explanation</b>	12	16	12	16	12	16	16	16	20	20

**Table 4.1. Accessibility System Supports Adjustments**

<b>Accessibility Support or Feature</b>	<b>Description</b>	<b>Expected Delivery</b>
Provide guidance regarding use of multiple accommodations by a single examinee	Identify and share resources on use of multiple accommodation supports used simultaneously. Bundling of supports is a common, frequently used practice. This enhancement of the ACT Aspire Accessibility User Guide will provide guidance discussion and examples of requirements, best practices and cautions to consider when a user seeks to simultaneously use multiple accessibility and accommodation supports.	Spring 2017
Expand Spanish Translation Forms to all grades 3–HS	Currently only grades 3–6 have Spanish translation of items available (in Writing, Math and Science tests where construct permits). Now all grades will have this direct linguistic EL support. (North American Spanish)	2015–2016
Activate Color Contrast Palettes	Removes a construct irrelevant barrier for examinees with certain low vision or color blindness concerns	2015–2016
Activate digital CBT highlighter tool	Connected to color contrast technical issue described above	2015–2016
Custom masking	Allows students to mask parts of text and show only, say, a few lines of the passage, to help student focus	2015–2016
Online embedded American Sign Language video	Removes construct-irrelevant barriers for students whose primary language is ASL	2015–2016
Online interactive screen-reader compatibility	Provides a fully independent test administration for students with blindness; also avoids security concerns with current pdf approach.	2015–2016
Map accessibility profiles across item interaction formats and task models for 7 learner populations	Provides shared structure and data around accessibility issues	2015–2016
Build and Launch Embedded ASL Video	Working with GAAP national research group on best practice procedure for development; this support will remove construct irrelevant barriers for examinees with deafness whose primary language is ASL	Spring 2017
Build and Launch Interactive CBT Screen Reader Compatibility	This will embed appropriate semantic tagging structures into interactive CBT content to provide a fully independent test administration digital interface for users with blindness. Launch of this support removes the need for these examinees to have a personal testing assistant present throughout testing to provide navigation and response support. It also overcomes important security concerns with providing pdf-based screen reader support.	Spring 2017



## *MCAS Response to HQAP High School Study*

Our goal as a Commonwealth is to ensure that every Massachusetts student is prepared to succeed in postsecondary education and compete in the global economy. We have been administering annual assessments in Massachusetts since 1998 as our way of holding ourselves accountable for our progress toward this goal. The Massachusetts Comprehensive Assessment System (MCAS) tests are generally considered the gold standard of state assessments. They hold students to high expectations—in most cases, equivalent to the proficiency standard on the National Assessment of Educational Progress (NAEP)—and use a variety of question formats to ensure that we assess the full range of student abilities. Over the years we have refined the assessments to adapt to changes in the curriculum frameworks, most notably the incorporation of the Common Core State Standards into our 2010 frameworks, and to improve the quality of the assessment over time.

Our students and educators have accomplished incredible things under this system. Massachusetts' NAEP scores have moved from middle of the pack to leading the nation, and our students have scored well on international assessments. We have also made substantial progress toward closing the proficiency gaps between student subgroups, and we have dramatically reduced our dropout rate and increased our cohort graduation rate. That success would not have been possible without a high quality assessment providing feedback on student, school, district, and state achievement and progress.

The Massachusetts Comprehensive Assessment Systems was a terrific twentieth-century assessment—but it has reached a point of diminishing returns. In 2015, MCAS was administered for the eighteenth year. We have a better understanding now than we did a decade or two ago about learning progression in mathematics, text complexity and the interplay of reading and writing, and the academic expectations of higher education and employers. And we now know that nearly one-third of our public high school students who go on to enroll in Massachusetts public colleges take at least one remedial course in their first semester, suggesting that the curriculum and assessments they have experienced have not adequately prepared them for the world beyond high school. Indeed, MCAS was never designed to be an indicator of college and career readiness. We joined the Partnership for Assessment of Readiness for College and Careers (PARCC) consortium specifically in order to partner with other states in developing an assessment that is more closely aligned to these expectations.

Thus, we were not surprised by this report's conclusion that the MCAS does not always measure well what's most important today. This report also confirms that in many ways, PARCC sets a higher bar than MCAS for student performance. This is particularly true as students move up the grades into middle and high school. This higher bar is not simply about being harder: PARCC provides more opportunities for critical thinking, applying knowledge, research, and making connections between reading and writing. More and more schools have upgraded curriculum and instruction to align with our 2010 frameworks. While we adjusted MCAS to test those frameworks, PARCC was built around them. Classroom instruction is now increasingly focused on the knowledge and skills in the frameworks, rather than how to pass a test.

We are proud of what we have accomplished in Massachusetts in the nearly two decades that we have been administering the MCAS. Now that we have the benefit of that experience and have revised our curriculum frameworks to reflect our upgraded learning expectations, it is time to upgrade our assessments too. Our state Board of Elementary and Secondary Education voted in November 2015 to do exactly that.

**2015–16 Test Program Changes:**

Over the next few years, we will transition to a new statewide assessment system that will take much of what this report identifies as the strengths of PARCC—high-quality content aligned strongly to college and career ready standards—and combine it with elements of MCAS in the context of a Massachusetts-specific governance system that will allow us to set our own policies on test content, administration, and reporting. With this approach, we will continue to benefit from a high-quality, next-generation assessment while ensuring that the test will reflect the Commonwealth’s unique needs and concerns. Most importantly, our students will be better prepared for success after high school—our ultimate goal.

## PARCC Response to HQAP High School Study

### ELA/L Rating on Speaking & Listening

Group Summary Statement:

- *None of the items assessed speaking and listening skills required for college and career readiness. This criterion was not included when establishing the composite Content rating.*

PARCC Response:

- The PARCC assessment measures many aspects that are key to the Speaking and Listening standards. PARCC uses multimedia texts to measure comprehension for all students taking its tests online (providing students with opportunities to demonstrate strengths and needs in comprehending audio and audiovisual texts). The CCSS build coherence across the ELA strands and identify similar skills built into both the reading comprehension standards (standards RI 7 and RL 7) and the listening standards. PARCC chose to report students' speaking and listening performance in relation to the reading standards.

The PARCC assessment system includes a robust set of Speaking and Listening tools. All schools administering PARCC in 2015-2016 have access to a comprehensive set of formative assessments and instructional tools to support educators, parents, and students in better understanding students' strengths and needs in speaking and listening. Further information about the PARCC Speaking and Listening tools can be found on PARCC's Partnership Resource Center:

<https://prc.parcconline.org/library/speaking-and-listening-overview>

### ELA/L Rating on Text quality and types

Group Summary Statement:

- *The texts in this assessment were of high quality and used open sources, but reviewers judged them to be overly rigorous. A larger range of text structure and purposes were needed to meet the Criteria; specifically, reviewers found that less than half of the passages on this assessment were informational, while the Criteria recommended about two-thirds of the texts be informational. Of the passages that were informational, the large majority was expository; however, the informational texts did not represent literary nonfiction, history/social science, and science/technical and, thus, the three categories were not evenly split, as recommended by the Criteria.*

*It should be noted there are typically only a limited number of passages that can be included on any given assessment, thus, the methodology's recommended distribution of passage types could be influenced greatly by a single discrepancy that might result in a different (lower or higher) rating.*

## PARCC Response:

The PARCC assessment measures both general informational reading (ELA) and the literacy standards (focus on general information, social science and history texts, and science/technical texts). We achieved the balance of informational and literary texts called for in the standards by considering numerous factors (e.g. the length of texts, the length of time students will spend engaged in reading and responding to each type of text, the number of items and points devoted to each type of text, the range of different genres for each type of text and potential ways to vertically articulate this range, and how to ensure the range of texts called for in reading standard 10 are demonstrated), as opposed to a raw count on the number of texts. For grades 6-11, where we measure literacy standards, as well as the ELA standards, it was important to ensure we had sufficient texts in history/social science and science/technical texts to allow for measurement of those standards.

### ELA/L Rating on Cognitive Demand

#### Group Summary Statement:

- *Reviewers found the distribution of cognitive demand on this assessment only partially matched the distribution of cognitive demand of the standards. Reviewers felt there were a lot of items at high DOK levels and not enough items at the lower levels, making it difficult to adequately assess the full range of student abilities.*

*It should be noted that assessment programs implement different cognitive complexity frameworks that might impact this rating, especially those programs that do not use Webb's DOK methodology to determine cognitive demand, as used in this study. PARCC uses the Cognitive Complexity Framework.*

#### PARCC Response:

- It is important to note that students who meet Level 1/Level 2 Depth of Knowledge (DOK) for items situated at higher DOK levels are given partial credit points for demonstrating skills that require lower cognitive complexity. Reviewers did not consider the possibility that scoring, rather than adding more Level 1 or Level 2 items, could allow for the balance of item complexities. For more information on the PARCC scoring rubrics and to view released items, visit: <https://prc.parcconline.org/assessments/parcc-released-items>

The PARCC assessment uses a cognitive complexity framework that was developed by the PARCC consortium to more accurately reflect the demands of the CCSS. This framework received recognition from AERA [2014 Outstanding Contribution to Practice in Cognition and Assessment award]. An article detailing the innovations of this framework and potential next steps in research around cognitive complexity has been published in a new book titled *The Next Generation of Testing: Common Core Standards, Smarter-Balanced, PARCC, and the Nationwide Testing Movement* (IAP—Information Age Publishing): <http://www.infoagepub.com/products/The-Next-Generation-of-Testing>)

## *Smarter Balanced Response to HQAP High School Study*

January 12, 2016

On behalf of the Smarter Balanced Assessment Consortium, thank you for your comprehensive study of our summative high school test's alignment to the Common Core State Standards and of our extensive accessibility features and accommodations.

This report confirms that Smarter Balanced is a high quality end-of year test that assesses students on what they are learning in their classrooms: the Common Core State Standards. We are pleased that reviewers overwhelmingly rated Smarter Balanced as having an “excellent match” in math and English-language arts for both content alignment and measuring the depth of content. It represents another seal of approval echoed by the nation’s top teachers in a recent review sponsored by the National Network of State Teachers of the Year.

It is because of the superior quality of Smarter Balanced that more than 200 colleges and universities in eight states representing public higher education systems and select private institutions use the assessment system to help students further their postsecondary education. These institutions have agreed to use scores from the Smarter Balanced high school assessments as evidence that students are ready for credit-bearing courses and can bypass non-credit, developmental courses. This will be welcome news to students, parents, and policy makers who know remediation costs students time and money, and students who must take these courses are far less likely to ultimately earn a degree.

The Smarter Balanced end of year test includes a comprehensive set of universal tools, designated supports, and accommodations that are unprecedented in assessment. Utilizing the principles of universal design, these accessibility resources include Braille, stacked Spanish translations, videos in American Sign Language, glossaries provided in 10 languages and several dialects, as well as translated test directions in 19 languages. Each of these accessibility resources was built with students in mind and would be cost prohibitive for any state to create on their own. While the report does not rate these extensive resources, it is clear from the documentation that students who take Smarter Balanced have the tools they need to show what they know and can do.

Measuring college and career ready standards requires more than multiple-choice items. By giving students the opportunity to demonstrate their critical thinking and use their problem solving skills, Smarter Balanced provides students, parents and educators with accurate information on how well students are prepared for the rigors of college and the modern workplace. As noted by one of the state teachers of the year who reviewed the Smarter Balanced assessments, educators need this information to help schools improve:

“I can’t help but think of students who went into college believing they were prepared only to find they needed remedial education to learn what they should have in high school. Educators must demand that schools do a better job preparing students and the new assessments are one tool to do that.”

Kristie Martorelli, 2012 Arizona Teacher of the Year

In every case, those who evaluate Smarter Balanced see the quality of the assessment system. We echo one reviewer's comments in this report, "This was so beneficial. My comfort level giving SBAC and teaching the CCSS is hugely increased." As educators become more knowledgeable about Smarter Balanced, we are confident our assessment system will continue to be recognized as a historic and groundbreaking system to improve teaching and learning.

Sincerely,



Tony Alpert  
Executive Director  
Smarter Balanced Assessment Consortium

## Appendix L: ACT Aspire Accessibility Summary Statements

### ACCESSIBILITY – ACT ASPIRE ELA/LITERACY AND MATHEMATICS SUMMATIVE ASSESSMENTS<sup>a</sup>

#### OVERALL SUMMARY

**Strengths:** The ACT Aspire summative assessments are administered online or as a paper version, by each state’s choice. The program provides a range of accessibility features and accommodations (e.g., eliminating irrelevant language demand, color contrast, limiting motor load, avoiding extraneous graphics), with similar accessibility features and accommodations offered for the paper-based and online assessments. Documentation includes a rationale for how each feature or accommodation supports valid score interpretations, when each may be used, and instructions for administration. ACT Aspire demonstrates strong adherence to universal design principles in its development of the assessed content areas. The program presented information about the types of accommodations available (see ACT Aspire link below) and the type of student who might benefit from each based on best practices and research.

**Areas for Improvement:** It was unclear how the program used information about the types of accommodations available and the type of student who might benefit from each when developing items and assembling forms. Also, the program’s implementation of its universal design principles may not have been fully realized during item development and form assembly. For example, reviewers found documentation that indicated the program would provide multiple accommodations but within the documentation provided they were unable to find information about how the program would manage providing multiple accommodations for a single student.

**Link to Documentation:** The list of accessibility features and accommodations offered by ACT Aspire can be found at [http://www.discoveractaspire.org/pdf/2014\\_actaspire\\_Accessibility\\_UserGuide2.0d.pdf](http://www.discoveractaspire.org/pdf/2014_actaspire_Accessibility_UserGuide2.0d.pdf).

Criterion	Group Summary Statement
<p>A.5: Providing accessibility to all students, including English learners and students with disabilities.</p>	<p><b>Construct Validity.</b> Reviewers found the accommodations and accessibility features offered for the ACT Aspire ELA/literacy and mathematics summative assessments were appropriate and based on research. The ACT Aspire documentation provided a rationale for the accommodations as well as definitions for the constructs assessed. Reviewers had some concerns about the construct validity for allowable accommodations and modifications; they could not find sufficient information about how the accommodations offered to students ensured the item construct was not impacted. In addition, documentation was limited in how test quality will be monitored or improved over time.</p> <p><b>Item Development.</b> Procedures for developing items for the ACT Aspire summative assessments were included in the program documentation; however, documentation indicated the program employed limited review of empirical data regarding accessibility features and accommodations at key points in the item development process. For example, cognitive labs were limited and did not include think-aloud or similar testing for students using accessibility features or accommodations. Documentation included information about generating the test forms, which reflected the principles of universal design and sound testing practice, as recommended by the criteria.</p> <p><b>Test Assembly Procedures.</b> Procedures for assigning students to forms and accommodations were included in the program documentation. However, reviewers were unable to locate documentation for assigning or managing multiple accommodations for either ELs or SWDs, as recommended by the criteria.</p> <p><b>Accommodation Policy.</b> The accommodations offered by the ACT Aspire ELA/literacy and mathematics summative assessments were documented, including a rationale for how each accommodation supports valid score interpretations, when each accommodation may be used, and instructions to administer each accommodation. Reviewers were not able to locate information related to how test quality is monitored or improved, as recommended by the criteria.</p>

<sup>a</sup> Only qualitative statements are provided for the accessibility criterion.

## Appendix M: MCAS Accessibility Summary Statements

### ACCESSIBILITY – MCAS ELA/LITERACY AND MATHEMATICS SUMMATIVE ASSESSMENTS<sup>a</sup>

#### OVERALL SUMMARY

**Strengths:** The MCAS summative assessments are paper-based. The program offers standard accommodations that change the routine conditions under which a student takes the MCAS (e.g., frequent breaks, unlimited time, magnification, small group) and nonstandard accommodations (modifications) that change a portion of what the test is intended to measure (e.g., read aloud or scribe in ELA, calculator or non-calculator portions of mathematics). These accommodations are provided to students with disabilities as determined by their Individualized Education Plan or 504 Plan and in accordance with the state’s participation guidelines. In general, reviewers judged the accommodations and accessibility features offered by MCAS for its summative assessments to be reasonable. MCAS documentation reflected the program’s efforts to consider universal design.

**Areas for Improvement:** There were limited accommodations indicated specifically for ELs. (MCAS provided the *Requirements for the Participation of English Language Learners* document after this study was completed that addressed, at least in part, deficiencies that reviewers found.) Although reviewers judged the accommodations and accessibility features offered by MCAS to be reasonable, they also thought they were limited and did not maintain pace with the field. Currently, the program’s use of universal design was perceived to be limited (based on the narrow populations considered and the limited feedback obtained during item development and bias reviews). The program offers a limited scope of accessibility features for some items and certain accommodations appear to introduce the opportunity for errors because student responses need to be transposed or items had to be skipped. Reviewers did not find a strong connection between research and the accommodations that MCAS made available in the provided documentation. After the study was completed, MCAS clarified that their manuals were written to be accessible and useable by the field; therefore, much of the research studies and policy explanations were not included in them. It is possible these additional documents might have addressed deficiencies that reviewers noted.

**Link to Documentation:** For this study, we used the Accessibility Manual for 2014 assessment to coincide with the materials provided for this evaluation. A full list of accessibility features and accommodations currently offered by the 2015-2016 MCAS is available at their website, <http://www.doe.mass.edu/mcas/participation/ell.pdf> and <http://www.doe.mass.edu/mcas/participation/sped.pdf>.

Criterion	Group Summary Statement
<p>A.5: Providing accessibility to all students, including English learners and students with disabilities.</p>	<p><b>Construct Validity.</b> As recommended by the criteria, reviewers found that the MCAS program documentation did include potential threats to validity through the use of universal design. However, some of the definitions for the assessed constructs were confusing (they did not clearly define a threat that might require accommodations or accessibility features) and certain definitions could be unnecessarily limiting, especially for students with disabilities.</p> <p><b>Item Development.</b> Reviewers were not able to find evidence that MCAS reviewed empirical data regarding accessibility at key points in the item development process, such as from cognitive labs or other focused try-outs. Also missing or not found were the processes that MCAS uses to support continual improvement of accommodations and accessibility features offered, especially those for English learners.</p> <p><b>Test Assembly Procedures.</b> Reviewers found limited documentation about how test forms as assembled and administered for students whose accommodations affect the selection of content (e.g., dual translation forms, Braille forms) and for how to assign appropriate accommodations and accessibility features to them. Also not found as how the program detects or corrects unwanted interactions between multiple accommodations or accessibility features.</p> <p><b>Accommodation Policy.</b> As recommended by the criteria, reviewers found evidence that the accommodations provided to ELs and SWDs were sufficient to support valid score interpretations. However, a number of accessibility needs were not addressed for ELs (i.e., most of the information focused on SWDs with no separate mention of ELs) and some of the accommodations provided to SWDs were unnecessarily restrictive. For example, high school students were permitted to use accommodations not allowed for younger students even though those allowances did not appear to interfere with construct validity and could be beneficial to SWDs.</p>

<sup>a</sup> Only qualitative statements are provided for the accessibility criterion.



## Appendix N: PARCC Accessibility Summary Statements

### ACCESSIBILITY – PARCC ELA/LITERACY AND MATHEMATICS SUMMATIVE ASSESSMENTS<sup>a</sup>

#### OVERALL SUMMARY

**Strengths:** The PARCC summative assessments are administered online and offer paper-based assessments for students, as appropriate. The program incorporates accessibility features that are available to all students (e.g., color contrast, eliminate answer choices, highlight tool, pop-up glossary) and offers several test administration considerations for any student (e.g., small group testing, separate location, adaptive and specialized equipment or furniture), as determined by school-based teams. The program also offers a wide range of accommodations for SWDs (e.g., assistive technology, screen reader, Braille note-taker, word prediction external device, extended time) and ELs (e.g., word-to-word dictionary, speech-to-text for mathematics, general directions provided in a student’s native language, text-to-speech for the mathematics assessment in Spanish). PARCC was viewed favorably for its sensitivity to the design of item types that reflect individual needs of students with disabilities, and for its strong research base and inclusion of existing research on ELs. Reviewers found the accommodations offered by PARCC to be valid and appropriate based on current research.

**Areas for Improvement:** Based on the information reviewed during the evaluation, reviewers were unable to locate information about the research needed to determine whether the accessibility features and accommodations that are offered by the program alter the constructs measured in its assessments. Specifically, reviewers noted that clearer documentation may be needed regarding how PARCC administers multiple features simultaneously and the implications of how multiple accessibility features impact student performance. After the workshop, PARCC provided information about how they conduct trials and customer acceptance testing to ensure multiple features and embedded accommodations are properly working that might have addressed deficiencies that reviewers found.

**Link to Documentation:** A full list of accessibility features and accommodations offered by PARCC is available on their website,

[http://www.parcconline.org/images/Assessments/Accessibility/PARCC\\_Accessibility\\_Features\\_Accommodations\\_Manual\\_v.6\\_01\\_body\\_appendices.pdf](http://www.parcconline.org/images/Assessments/Accessibility/PARCC_Accessibility_Features_Accommodations_Manual_v.6_01_body_appendices.pdf)

Criterion	Group Summary Statement
<p>A.5: Providing accessibility to all students, including English learners and students with disabilities.</p>	<p><b>Construct Validity.</b> The constructs assessed on the PARCC ELA/literacy and mathematics summative assessments were defined with sufficient clarity; however, reviewers could not find rationales for the construct definitions that incorporated available research. Information about providing modifications and their restrictions was included, but more information was needed about how threats to validity are addressed through universal design, accommodations, and accessibility features. As recommended by the criteria, PARCC includes a process for improving its support of validity regarding accessibility and accommodations.</p> <p><b>Item Development.</b> PARCC’s item development procedures regarding accessibility were perceived to build on the construct definitions and be sufficiently systematic such that reviewers were able to verify the claims made conceptually and empirically about the constructs. Documentation on item development included a number of expert reviews as well as a plan to use those expert reviews to attend to potential challenges (e.g., particular disability, socioeconomic condition).</p> <p><b>Test Assembly Procedures.</b> PARCC documentation included procedures for how forms are to be assembled and administered to students whose accommodations affect the selection of the content as well as how appropriate accommodations should be assigned to individual students. However, although recommended by the criteria, reviewers could not find procedures for detecting and correcting unwanted interaction among multiple accommodations and access features. PARCC’s field test results provided evidence of the program’s intent to monitor and improve the quality of its test assembly procedures that consider accessibility; however, the information was not disaggregated by disability type. Reviewers also were unable to find information that described possible interactions between multiple accommodations offered.</p> <p><b>Accommodation Policy.</b> Reviewers found that PARCC addressed the accessibility and accommodation needs of the vast majority of students. As recommended by the criteria, PARCC documentation indicated that accommodations were based on research sufficient to support valid score interpretation, credible use of scores, and legal defensibility. Reviewers noted that a line of validity research is planned that will include the impact of accommodations on ELs and SWDs as well as ELs with disabilities.</p>

<sup>a</sup> Only qualitative statements are provided for the accessibility criterion.

## Appendix O: Smarter Balanced Accessibility Summary Statements

### ACCESSIBILITY – SMARTER BALANCED ELA/LITERACY AND MATHEMATICS SUMMATIVE ASSESSMENTS<sup>a</sup>

#### OVERALL SUMMARY

**Strengths:** The Smarter Balanced summative assessments are administered online as adaptive tests as well as via paper-based versions. The program provides a range of accessibility resources: universal tools, designated supports, and accommodations. Depending on preference, students can select a number of universal tools that are embedded (e.g., digital notepad, highlighter, zoom, English glossary) or non-embedded (e.g., protractor, scratch paper, thesaurus, English glossary) within the assessment. The program also offers a number of designated supports to all students for whom the need has been indicated by an educator or team of educators. The designated supports can be embedded (e.g., color contrast, magnification, translations for the online version, translated glossary) or non-embedded (e.g., color contrast, separate setting, translations for the paper or online versions, translated glossary). For students with documented Individualized Education Plans or 504 Plans, several embedded accommodations are available (i.e., American Sign Language, Braille, closed captioning, and text-to-speech) and several non-embedded accommodations (e.g., abacus, read aloud, scribe, and speech-to-text) are offered. The program has specific guidelines for accessibility for ELs that highlight using clear and accessible language when developing items. Smarter Balanced’s use of universal design and evidence-based design were described well. The program also appropriately suggests usability guidance to help educators support determinations of how different accommodations, designated supports and universal tools might interact.

**Areas for Improvement:** The program’s item development procedures incorporated accommodations and accessibility features from conception, which is consistent with the criteria. However, decision making rules were judged to be overly complicated and challenging for educators to apply. For SWDs, certain guidelines were judged to be overly prescriptive when there did not seem to be a reason for such strict guidance. After the workshop, Smarter Balanced highlighted the usability guidance that helps educators support determinations of appropriate accommodations, designated supports and/or universal tools and how they might interact in the *Individual Student Assessment Accessibility Profile* (ISAAP) documentation. This information may have addressed, at least in part, deficiencies that reviewers noted.

**Link to Documentation:** A full list of accessibility features and accommodations offered by Smarter Balanced is available on their website, [http://www.smarterbalanced.org/wordpress/wp-content/uploads/2014/08/SmarterBalanced\\_Guidelines.pdf](http://www.smarterbalanced.org/wordpress/wp-content/uploads/2014/08/SmarterBalanced_Guidelines.pdf).

The ISAAP interactive module, directions for using the ISAAP tool, and the web-based tool are available online:

- <https://www.smarterbalancedlibrary.org/content/introduction-individual-student-assessment-accessibility-profile-isaap-updated?key=4089340e54d1b0e96ae053fee694f178>
- [http://52.11.155.96/static/isaap/ISAAP-Web-Tool-Version%201-Instructions\\_081322015.pdf](http://52.11.155.96/static/isaap/ISAAP-Web-Tool-Version%201-Instructions_081322015.pdf)
- <http://52.11.155.96/static/isaap/index.html>

Criterion	Group Summary Statement
<p>A.5: Providing accessibility to all students, including English learners and students with disabilities.</p>	<p><b>Construct Validity.</b> Smarter Balanced program documentation provided clear definitions of the constructs assessed, which easily distinguished between construct relevant and irrelevant variance.<sup>b</sup> Program documentation also provided rationales and incorporated current research into its construct definitions. The program defined threats to validity that might require accommodations or accessibility features. Information was included about how the program planned to improve its support of accessibility and accommodations; however, reviewers were unable to find details as to how the program plans to improve its support of accessibility and accommodations in the documents reviewed during the workshop. Smarter Balanced provided additional clarification for program improvement that is outlined in their <i>Usability, Accessibility, and Accommodations Guidelines</i> after the workshop. The information may have addressed, at least in part, deficiencies reviewers noted.</p>

ACCESSIBILITY – SMARTER BALANCED ELA/LITERACY AND MATHEMATICS SUMMATIVE ASSESSMENTS <sup>a</sup>	
	<p><b>Item Development.</b> Reviewers judged that items were developed in a manner consistent with the principles of universal design and evidence-based design. Smarter Balanced item development procedures were perceived to build on definitions of the constructs, and the accommodations and accessibility features faithfully maintained those constructs during assessment. Documentation described appropriate expert and empirical data reviews regarding accessibility at key points in the item development process. Documentation included instructions for identifying how and when accommodations and accessibility features may be administered.</p> <p><b>Test Assembly Procedures.</b> Reviewers found limited documentation about how test forms are assembled and administered for students whose accommodations affect the selection of content; however, Smarter Balanced commented they highlighted information about detailed simulations for specific accessibility including Braille and ASL. The program’s decision making rules were judged to be overly complicated for educators to easily apply and the guidance not sufficiently explicit for schools to implement. During the workshop, reviewers were unable to find information about how the program plans to collect, analyze, and act on information to monitor and improve the quality of its test assembly procedures that consider accessibility; however, Smarter Balanced responded that their simulations are a systematic method of identifying areas that need improvement and derive item writing plans to augment the item pool for subsequent years. This information and accompanying documentation might have addressed deficiencies that reviewers found</p>
	<p><b>Accommodation Policy.</b> Documentation indicated that accommodation and accessibility features were based on research sufficient to support valid score interpretations, credible use of scores, and legal defensibility. Reviewers found the program addressed all major accessibility needs, with the range of accessibility tools diverse and extensive for both ELs and SWDs. While there was an accurate list of accommodations, only limited information was found regarding variations in the frequency of administration due to policy or technical information about the impact on validity and comparability of score interpretations due to such policy. Smarter Balanced commented that they will collect and analyze these data as states make them available.</p>

<sup>a</sup> Only qualitative statements are provided for the accessibility criterion.

<sup>b</sup> Construct relevance refers to the factors that are related to what the assessment is intended to measure while construct irrelevance refers to the factors that are not related to what is being measured.