GUIDE TO EVALUATING ASSESSMENTS USING THE CCSSO CRITERIA FOR HIGH QUALITY ASSESSMENTS:

Focus on Test Content

MARCH 2016

Center for Assessment

TABLE OF CONTENTS

TABLE OF CONTENTS

TABLE OF CONTENTS *continued*

TABLE OF CONTENTS

INTRODUCTION

How can the quality of assessments of college and career ready standards be evaluated? Based on criteria established by the Council of Chief State School Officers (CCSSO), the National Center for the Improvement of Educational Assessment (Center for Assessment) developed a methodology focused on test content. This document provides a detailed description of the methodology and provides information about the development process

THE NEED

Reports of student achievement and growth are valued by students, parents, educators, policymakers, and the public. In particular, as states adopt standards to help students be ready for college and careers, many new assessments have been developed that intend to assess and report on students' progress toward these learning goals. College and career ready standards challenge students "to develop a deeper understanding of the subject College matter, learn how to think critically, and apply what they are learning to the real world." (CCSSO, States' commitment to high-quality assessments aligned to college- and career-readiness, 2013, p. 1)   To realize the promise of new standards and to inform better teaching and learning, state assessments must be high quality and must match the standards in rigor and depth. How can state officials responsible for administering state assessments and educators, policymakers, and others desirous to interpret and use assessment results identify assessments that meet demanding criteria of quality?

To address that question, several coordinated efforts are required.

NECESSARY PARTS FOR AN EVALUATION

There are four essential components needed in order to know to what degree an assessment meets criteria of quality:

1. Criteria that delineate essential aspects of quality
2. A methodology for evaluating the assessment in terms of those quality criteria
3. An evaluation study that implements the methodology in a credible way and reports the results in an understandable and useful form
4. An assessment to be evaluated with supporting documentation

The Council of Chief State School Officers (CCSSO), the member organization representing the head state education officers developed in 2014 *Criteria for Procuring and Evaluating High Quality Assessments* (referred to hereafter as the *CCSSO Criteria*). The *CCSSO Criteria* are summarized below.

The Center for Assessment developed a **methodology** for applying the *CCSSO Criteria* to assessments that might be used by state departments of education for summative purposes. The methodology involves the examination of actual assessment items as well as key assessment program documentation (e.g., test specifications) by panels of qualified experts. This document provides an overview of that methodology and specific methodological components in the Appendix.

The methodology must be implemented carefully by a group of evaluators, in an **evaluation study** organized by an implementer who gathers and organizes the necessary documentation, trains and organizes the evaluators, provides practical ways to conduct the evaluation ensuring appropriate confidentiality and security of materials, monitors accurate reporting of results, publishes the report, and so on. This document does not include any results from an evaluation study. It is expected that results from evaluation studies using the Center's methodology will be published by the organization responsible for implementing and/or sponsoring the evaluation study. In particular, two such evaluation studies using the Test Content methodology are expected to provide examples of the implementation of the methodology, as well as evaluation results for several assessments. This is discussed further in the section on *Development of the Methodology*.

THE CCSSO *CRITERIA*

CCSSO published its *Criteria for Procuring and Evaluating High Quality Assessments* in March 2014. The CCSSO *Criteria* were "intended to be a useful resource" for "states to consider as they develop procurements and evaluation options for high-quality state summative assessments aligned to college- and career-readiness standards." In particular, the *Criteria* were "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" (p. 1). The CCSSO *Criteria* are organized under six main topics.

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

The CCSSO *Criteria* document explicates each of these main topics into criteria. For example, the topic, B. *Align to Standards – English Language Arts/Literacy* is expanded into nine criteria, B.1-B.9. The *CCSSO Criteria* document also includes more detailed descriptions of criteria and sample evidence. (A link to the CCSSO *Criteria* document is provided under *Other Resources* at the end of this document.)

The Center for Assessment has translated the CCSSO *Criteria* into more specific rubrics and scoring procedures to support practical and credible evaluation of the *Criteria*. To facilitate development of the evaluation methodology, the Center for Assessment partitioned the CCSSO *Criteria* into two logical sections: one dealing with Test Content and the other dealing with Test Characteristics. Test Content focuses on alignment to standards (Criteria B.1-B.9 and C.1-C.5), as well as on providing accessibility to all students (A.5) and transparency of test design and expectations (A.6). This document describes the methodology for evaluating Test Content. The Center for Assessment's methodology for evaluating Test Characteristics, which will address all the remaining CCSSO Criteria, is under development and expected to be available in 2016.

In response to questions identified when the Center for Assessment was developing the evaluation methodology, CCSSO developed more explicit guidance to supplement the *CCSSO Criteria*. Some of the notable additions are that CCSSO grouped the alignment *Criteria* into those dealing with *Content* and *Depth*. CCSSO also provided additional guidance on sufficiency of evidence and weighting of various criteria. This supplemental guidance from CCSSO is provided in the Appendix.

This methodology is the first attempt to operationalize the CCSSO Criteria and create a methodology suited to review of college and career ready standards. As future Implementers use the methodology to review a variety of assessments, it is likely that that they will identify ways the methodology could be improved. Thus, this methodology will be a living document and is likely to be enhanced in the future.

OUTCOMES OF AN EVALUATION OF TEST CONTENT

The primary outcomes of an evaluation of an assessment in terms of Test Content will be a profile of **ratings** and a corresponding set of **comments**. In addition, Test Content criteria will be grouped into two categories, "Content" and "Depth," which will also receive ratings. The ratings for each criterion and for Content and Depth will be Weak, Limited/Uneven, Good, or Excellent Match to the criterion. The results across the criteria can be interpreted as a profile of each assessment; no single overall rating will be generated.

In addition to the ratings, evaluators will produce Comments, which may include key information regarding the rationale for the rating or annotations of strengths and areas for improvement to help inform future development of the assessment program.

The CCSSO Criteria explicitly set a "high bar for quality" (p. 1) and the methodology seeks to take the same approach. A rating of "Excellent" on a criterion is intended to be a high bar which, if met, represents a more comprehensive measure of the knowledge and skills needed college and career readiness and/or a fairer way to assess students, particularly English Language Learners and students with disabilities, than is currently found in most state assessments. It is expected that most if not all assessments will have room for improvement in meeting the Criteria, and thus the Comments will provide feedback to inform assessment programs' continuous improvement efforts.

A sample Test Content summary report template is shown below. Each of the criteria will have a summary rating represented by the classifications of "Weak", "Limited/Uneven", "Good", and "Excellent". The filled-in circle with the corresponding color represents the summary of the decision after the reviews have been undertaken. Space for comments is provided. This represents a template of the summary; a full detailed report would follow this summary.

Results of Applying the CCSSO Criteria for High-Quality Assessments in Test Content

	Degree of Match with CCSSO Criteria			
	Weak	**Limited/Uneven**	**Good**	**Excellent**
A. Meet Overall Assessment Goals and Technical Quality – Accessibility & Transparency				
• A.5: Providing accessibility to all students, including English learners and students with disabilities (subset of the criterion)			●	
• A.6: Ensuring transparency of test design and expectations	●			
B. English Language Arts/Literacy				
I. Assesses the content most needed for College and Career Readiness *[[summary of rationale and other comments]]*		●		
• B.3: Requiring students to read closely and use evidence from texts	●			
• B.5: Assessing writing		●		
• B.6: Emphasizing vocabulary and language skills		●		
• B.7: Assessing research and inquiry			●	
• B.8: Assessing speaking and listening (optional)	●			
II. Assesses the depth that reflect the demands of College and Career Readiness *[[summary of rationale and other comments]]*			●	
• B.1: Assessing student reading and writing achievement in both ELA and literacy		●		
• B.2: Focusing on complexity of texts				●
• B.4: Requiring a range of cognitive demand			●	
• B.9: Ensuring high-quality items and a variety of item types				●

Results of Applying the CCSSO Criteria for High-Quality Assessments in Test Content (continued)

	Degree of Match with CCSSO Criteria			
	Weak	Limited/Uneven	Good	Excellent
C. Mathematics				
I. Assesses the content most needed for College and Career Readiness *[[summary of rationale and other comments]]*			●	
• C.1: Focusing strongly on the content most needed for success in later mathematics			●	
• C.2: Assessing a balance of concepts, procedures, and applications		●		
II. Assesses the depth that reflect the demands of College and Career Readiness *[[summary of rationale and other comments]]*			●	
• 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				●

EVIDENCE AND THE EVALUATION PROCESS

The Test Content methodology specifies what should be examined in the evaluation, who should conduct the evaluation, and how the evaluation should be conducted. These aspects are summarized below, and then a detailed example is provided to enable a reader to understand the basis for the evaluation of Test Content. In addition, the very specific Scoring Summaries for all the CCSSO sub-criteria used in evaluating Test Content are included in the Appendix.

General Description of Evidence and the Evaluation Process

What: The Test Content methodology is designed to answer the following question: “To what extent do assessments under review match the CCSSO Criteria relevant to Test Content?”. As the introduction to the CCSSO Criteria clarifies, these criteria “focus on the critical characteristics that should be met by high-quality assessments aligned to college- and career-readiness standards.” For literacy, this includes the careful examination of texts and meaningful work in reading and writing that centers on texts. For mathematics, this includes focusing on the mathematical content that matters most. In addition, for both literacy and mathematics, this includes a focus on ensuring that assessments are accessible for all students, including for students with disabilities and English Language Learners. As a result, the Test Content methodology does not prioritize one-to-one alignment of specific test questions (or items) to standards. Instead, the methodology operationalizes the CCSSO Criteria’s focus on higher-level features deemed most representative of the shifts required for assessments of college and career readiness.

The resulting methodology has numerous Criteria and each Criterion includes Sub-Criteria that represent important aspects of the Criteria to be considered. For example, Criterion B.1 “Assessing student reading and writing achievement in both ELA and literacy” has two sub-criteria dealing with a) the balance between types of texts (literary and informational) and b) the quality of the text passages used in the assessment.

In addition, the methodology requires reviewers to examine two types of evidence to inform their judgments regarding the quality of an assessment program and the extent to which each criterion is met: Outcomes and Generalizability. The first type of evidence comes from examination of assessment items and forms from actual operational tests. Evaluators consider assessment items that have been or will be operationally administered, presented as they were/will be administered (e.g., computer-administered items viewed on the computer platforms on which they were administered; paper-based items viewed in the actual test booklets or in a pdf). This provides direct evidence of what students will have experienced, and the resulting evidence is referred to as Outcomes evidence. The Generalizability evidence comes from examination of documentation provided by the assessment program (for example, test blueprints). This documentation provides evidence on what might be seen across all test forms reviewers could possibly see and helps reviewers determine whether results from the item and form review can likely be generalized across all forms the program might create.

This examination of both Outcomes and Generalizability evidence is useful because assessment programs often administer multiple forms of the same test. For example, a program may have 10 forms of a 4th grade math test with slightly different questions. And, for computer adaptive assessments, there will be a very high number of possible forms. However, in an evaluation of an assessment, it is generally only feasible for reviewers to examine one or two forms of each test in-depth. Looking at both Outcomes and Generalizability evidence allows reviewers to make judgments about the intent of the assessment program and whether the test forms reviewed are likely to be representative of all possible forms for a given test, as well as the quality of actual implementation in specific forms and items.

As noted above, the methodology is designed to be applied in the review of operational assessments. However, the methodology may usefully be adapted for other contexts. For example, for practical and logistical reasons Implementers might want to review sample items or released tests using this methodology. As another example, those wishing to procure assessments might use elements of the methodology in their procurement process to provide examples of the attributes testing programs should demonstrate or of the evidence that they should provide.

Who: The quality of an assessment evaluation depends in large part on the selection of qualified and experienced yet impartial reviewers. In recruiting and selecting reviewers, Implementers should look for reviewers with as many of the following qualifications as possible:

- Deep content knowledge -- reading, writing, or mathematics - for the specific grade span being reviewed
- Classroom content teaching experience; or experience as a district curriculum, reading or math leader (e.g., RtI supervisor, reading specialist, instructional coach), or special education supervisor
- Knowledge of/or familiarity with college and career ready standards in either mathematics or ELA/Literacy for at least one grade span
- Assessment or teaching experience working with English language learners and students with disabilities (e.g., having an understanding of Universal Design principles, linguistic features)
- General knowledge of large-scale assessment test specifications, test blueprints, and evidence-centered design principles
- Prior experience with assessment reviews
- Some familiarity with large-scale assessment test items, performance tasks, task templates that guide task design, and scoring rubrics/keys
- Understanding of importance of test security and willingness to keep confidential information
- Possesses the skills to work collaboratively (listen respectfully, honor divergent views, etc.)
- Ability and willingness to learn coding procedures for content and performance (rigor) analyses
- Ability to code ratings accurately according to directions; put aside personal opinions about any of the specific programs/products to be evaluated
- Ability to learn and enter rating information accurately in supplied computer software, if used

Each review panel should represent a range of these characteristics in order to provide an appropriate balance of expertise on each panel. Different panels for each content area (ELA and Mathematics) and for each grade level are generally preferable. In addition, if an Implementer is reviewing multiple assessment programs, there may need to be more than one panel per content and grade level so that no reviewer is asked to review an unreasonable number of tests and associated documentation in the time allotted. A jigsaw panel design is one way Implementers may choose to address this issue. The evaluations of Criteria B (ELA) and C (Mathematics) should be conducted by panels of 4-8 evaluators, although the Generalizability review of test documentation can be done by a smaller sub-panel composed of those with experience reviewing technical documentation.

For the accessibility review (see *Evaluation of Accessibility* for more), the panel may overlap with the panel evaluating the other aspects of Test Content, or it may be a separate panel that focuses only on test accessibility. A typical accessibility panel would consist of at least 3-4 persons who together have appropriate expertise. Typical areas of expertise would include the construct being assessed (e.g., reading), accommodation needs of special populations (e.g., English learners, students with disabilities), and the accommodations offered by the assessment program (e.g., technology-based accommodations). Because different issues arise for each content discipline, evaluators should consider disciplines separately (e.g., English language arts and mathematics); some members of the evaluation panel might need to be different to reflect the necessary disciplinary expertise.

Study Implementers are responsible for ensuring the evaluators are able to do what they are required to do to produce accurate ratings and comments. Accomplishing this should typically involve training on the specific procedures and materials of the evaluation study, as well as some type of monitoring that the evaluators can apply the training in following the procedures and making accurate judgments.

How: The test content evaluation methodology includes multiple steps in evaluating the assessments against the CCSSO criteria. Following training and calibration, reviewers first independently examine test items and passages and rate them on a series of criteria. Second, these reviewers use their item-level ratings to reach form-level ratings for each test form on each criterion. Third, reviewers engage in a process of discussion and consensus building, to move from reviewer-level results of a single test form to panel-level results across a testing program. In this process, they may draw on evidence from the Generalizability review of documentation to adjust their ratings. Whenever possible, the group ratings and statements should indicate consensus, but minority viewpoints may be expressed and recorded in the comments. Individual reviews are conducted for each sub-criterion, individual and group evaluations take place for Sub-criteria (e.g., B.1.1) and for each CCSSO Criterion (e.g., B.1), culminating in final group Content and Depth ratings. The pattern takes advantage of independent expert judgment and group discussion by expert judges evaluate complex and interacting dimensions. An illustration of this process is provided below.

Detailed Example of Evidence and Evaluation Process

The Test Content methodology identifies particular aspects to be evaluated associated with each CCSSO criterion, and provides a process and guidance for doing so. The set of guidance for an element is referred to as the “Scoring Summary” for that element. There is a Scoring Summary for each sub-criterion, criterion, and the content/depth aspects.

An example Scoring Summary for a sub-criterion is shown on the next page and described below. The Scoring Summary includes:

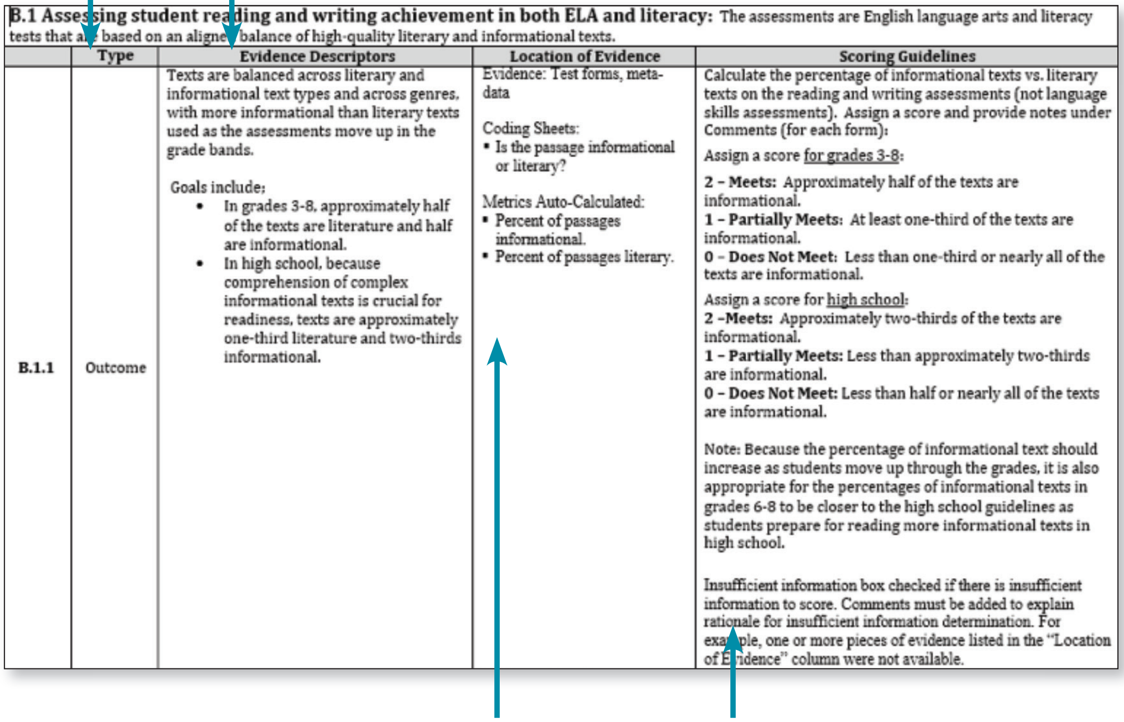
- **CCSSO Criterion** to be evaluated
- **Sub-Criterion** to be evaluated
- **Evidence Descriptors:** Description of the characteristics of the sub-criterion, and guidelines for what is acceptable evidence
- **Type of Evidence:** Whether the evidence is derived from examining operational test items/forms[1] (Outcomes) or from assessment program documentation (Generalizability).
- **Evidence:** Identifies the evidence that is provided by the assessment program and is to be examined by the evaluators. Also identifies the evidence that is produced by the evaluators in terms of coding and/or metrics that can be automatically generated based on the evaluators’ codings.
- **Scoring Guidelines:** Provides a rubric to guide evaluators in assigning a score/rating. Evaluators are also directed to provide appropriate comments.

Each of these aspects of the Scoring Summary is annotated in the example with a red arrow and explanatory comment in a box.

Type of evidence: Outcomes evidence is derived from examining test items/test forms; Generalizability evidence from the program documentation.

Evidence Descriptors: Description of the characteristics of the sub-criterion, and guidelines of what is acceptable evidence.

CCSSO Criterion number and description of what is to be evaluated



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
B.1.1	Outcome	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. Goals include: • In grades 3-8, approximately half of the texts are literature and half are informational. • In high school, because comprehension of complex informational texts is crucial for readiness, texts are approximately one-third literature and two-thirds informational.	Evidence: Test forms, meta-data Coding Sheets: ▪ Is the passage informational or literary? Metrics Auto-Calculated: ▪ Percent of passages informational. ▪ Percent of passages literary.	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): Assign a score <u>for grades 3-8</u>: **2 – Meets:** Approximately half of the texts are informational. **1 – Partially Meets:** At least one-third of the texts are informational. **0 – Does Not Meet:** Less than one-third or nearly all of the texts are informational. Assign a score for <u>high school</u>: **2 –Meets:** Approximately two-thirds of the texts are informational. **1 – Partially Meets:** Less than approximately two-thirds are informational. **0 – Does Not Meet:** Less than half or nearly all of the texts are informational. 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. 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.

Sub-criterion of the CCSSO Criterion to be evaluated

Identifies the evidence that is provided by the assessment programs and examined by evaluators; the evidence that is produced by the evaluators in terms of coding, and what metrics are automatically calculated based on the evaluators' codings.

For B.1.1, the assessment program provides the test forms and meta-data regarding the text passages. The evaluator codes whether the text passage is informational text. The percentage of text passages that are informational is automatically calculated by the coding software.

These scoring guidelines provide a rubric for evaluators to assign scores/ratings. For B.1.1, a test form needs to have approximately half of the texts be informational texts in order to receive a "Meets" score.

Evaluators are directed to provide appropriate comments as well.

Although there are many aspects to be evaluated, the evaluation methodology follows a general pattern, where *individual* evaluators consider evidence followed by *group* discussion of evidence and ultimately a *group* rating. This general pattern of evaluation is described below, with short examples. The example shows various coding and rating forms that evaluators are required to fill in. The forms provide structure for the evaluators' work, and become the recorded evidence of their work.

1. For sub-criteria designated as Outcomes, individual evaluators review operational test items/forms and make a judgment about the evidence associated with a specified aspect. The judgment may be preceded by descriptive coding of aspects of the assessment.

For example, an evaluator is required to make a judgment about whether the proportion of informational versus literary texts is consistent with the CCSSO *Criteria* and guidance in the Scoring Summary. To evaluate the *proportion* of text types, the evaluator must first examine each of the text passages on a test form and code whether the passage is informational or literary text. For more generalizability across forms, the evaluator may do the same for a second test form.

EXAMPLE CODING FORM FOR BALANCE OF TEXT TYPES, B.1.1

B1 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.			
Passage Identifier	Balance of text types		
	Is the passage informational?		

Using a coding form such as the one above, the evaluator would enter the Passage identifier for each text passage on the test (or the Passage identifier may already have been entered). The evaluator decides whether the passage is informational text and enters a code for each text passage (i.e., "Y" or "N"). If the coding form is electronic, the codes may be entered using a drop-down menu, which facilitates greater accuracy of recording results.

The result of this passage-by-passage review is a list of text passages in the test form, with a code assigned for each one by the evaluator. The percentage (proportion) of tested passages that are informational is calculated. In the example, the results are automatically summarized by an electronic coding form.

This table will automatically populate as you work. When you have finished, transfer the appropriate values to the rubric and assign a score. If you are reviewing multiple forms, do not assign a score until all forms are reviewed.	
B1 Totals	
Balance of text types totals	
Number of informational passages	0
Percent of tested passages that are informational	#DIV/0!

Based on the summary of the evidence and the scoring rubric in the Scoring Summary, the evaluator would rate how well the proportion of literary texts observed on the test forms met the CCSSO *Criteria*. The evaluator assigns a score of 0, 1, or 2 for the evidence from the operational test forms. To determine this score, the reviewer may draw on a tentative score that can be automatically produced by the coding form based on the scoring guidelines and use his/her professional judgment to adjust the score as needed (providing comments to justify any changes).

In the example of Sub-Criterion B.1.1, dealing with balance of information and literary text types, for grades 3-8, the Scoring Summary provides Scoring Guidelines:

2 – **Meets:** Approximately half of the texts are informational.
1 – **Partially Meets:** At least one-third of the texts are informational.
0 – **Does Not Meet:** Less than one-third or nearly all of the texts are informational.

Note that the full Scoring Summary for B.1.1 (available in the Appendix) includes Scoring Guidelines for the proportion of information text for grades 9-12 as well.

2. For sub-criteria designated as Generalizability, evaluators follow a similar process except that the evaluators examine program documentation in relation to the *CCSSO Criteria* and Scoring Summaries. It is also appropriate for a subset of reviewers to conduct the Generalizability review separately, in which case the results would be fed into the process during the group discussion and rating stage.

    For example, an evaluator makes a judgment about whether the proportion of informational versus literary texts is consistent with Generalizability Sub-Criterion B.1.4. The evaluator examines the documentation provided by the assessment program (e.g., test blueprints or other documents) and determines the extent to which the distribution of the types of passages specified by the documentation is consistent with the sub-criterion and lends support (or not) to the test-form ratings. As another example, reviewers may each reach their 0, 1, 2 scores relying only on Outcomes data and only take Generalizability data into account when they are rolling up results during the group discussion and consensus phase.

3. Evaluators then come together in grade level and content area panels to discuss their individual ratings and the evidence and to determine a group rating. This happens at key points in the process, notably for sub-criteria (e.g., B.1.1, B.1.2), for each CCSSO criterion (e.g., B.1), and for the final Content and Depth ratings.

    For example, in assigning a group rating for Sub-Criterion B.1.1, evaluators draw on their individual ratings, consideration of the evidence, and the scoring rubric.

    Similarly, a group rating is determined for the other sub-criterion (e.g., B.1.2).

    Finally, to inform their rating of "Weak" to "Excellent" for this B.1 criterion, the evaluators consider evidence regarding all the sub-criteria related to this criterion, which includes evidence from both the Outcomes and Generalizability sub-criteria. If the program documentation indicates that the rating for a criterion would likely increase or decrease if more forms had been reviewed, the evaluators will determine whether to adjust the final criterion rating and, if so, state the rationale.

    This same process of group deliberation is followed regarding ratings of combinations of criteria for Content and Depth.

    This example has aspects that involve content expertise (e.g., knowing whether a text passage is literary or informational); some aspects are low inference/low expertise, such as calculating the percentage of passages on the test that are informational. Some aspects require both content and assessment expertise, such as how to apply the Scoring Guidelines when the number of passages is small—so one passage has a large effect on the proportion of informational texts, or how to interpret assessment blueprints when one passage is long and has several assessment items or may be intentionally balanced by two shorter passages.

4. Evaluators record Comments at each stage to document the rationale for their ratings. The final report contains comments that provides a summary of the evaluators' rationale, and may also include strengths and areas to improve.

Scoring Summaries

The complete set of Scoring Summaries is provided in the Appendix.

- **English Language Arts Scoring Summary** – The ELA scoring summary is a compact synopsis of the basis for the evaluation. The scoring summary addresses the nine CCSSO *Criteria* for ELA/Literacy as well as A.5 for accessibility (7 Subcritiera) and A.6 for transparency in the context of ELA assessment, focused on evidence from operational forms supporting a rating on Outcomes (20 sub-criteria) and on evidence from documentation supporting a rating on Generalizability (20 sub-criteria).

- **Mathematics Scoring Summary** – The mathematics Scoring Summary addresses the five CCSSO *Criteria* for mathematics as well as A.5 for accessibility (7 Subcritiera) and A.6 for transparency in the context of mathematics assessment, focused on evidence from Outcomes (6 sub-criteria) and on evidence from documentation supporting Generalizability (9 sub-criteria).

Evaluation of Cognitive Demand

The CCSSO *Criteria* ask that the distribution of cognitive demand for each grade level and content area be sufficient to assess the depth and complexity of the standards (Criteria B.4 and C.4). Determining whether these Criteria are met requires four main activities:

A. Coding the content standards to determine what the target distributions of cognitive demand ought to be;
B. Coding the assessment items to determine what the distribution of cognitive demand is for the assessment test form(s);
C. Evaluating the observed cognitive demand of the assessment items in relation to the target cognitive demand of the content standards;
D. Evaluating the intended cognitive demand of the assessment program, as specified in documentation such as test specifications, in relation to the target cognitive demand of the content standards.

Appendix B provides guidance on how to conduct these four activities.

Evaluation of Accessibility

The CCSSO *Criteria* include accessibility, which reflects a concern with fairness, one of the fundamental aspects of validity in testing (AERA, APA, & NCME, 2014). CCSSO's accessibility criterion encompasses what would be considered accommodations and also access features. In the Test Content methodology evaluators focus on the adequacy of documentation provided by the assessment program; evaluators evaluate a sample of items and associated documentation. (A more complete evaluation of the validity of the accessibility of the program's assessments may be conducted as part of the Test Characteristics evaluation using on data from operational administrations.)

The review of the accessibility sub-criteria (A.5.1 –A.5.4) follows the same pattern of individual and group evaluation of evidence and determination of ratings. For sub-criteria designated as Generalizability, reviewers examine program documentation, which may include such things as 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, and empirical evidence from item-tryouts, etc.

For sub-criteria designated as Outcomes, reviewers examine exemplar items in ELA and mathematics that provide concrete evidence to ground their understanding of the assessment program's handling of accommodations/access features in conjunction with the program's documentation. An Exemplar may be an assessment item with a specific 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. The reason to examine Exemplars is a practical one. Each item on a form may be available with many different access features and accommodations (e.g. large print, highlighting, braille, text to speech, dictionary, translation) and reviewing every item on each reviewed test form in every different accommodated version and for every access feature is unlikely to feasible.

More detailed guidance on evaluating accessibility is provided in Appendix C.

Form Selection and Additional Summary Data on Forms

Every assessment program will likely have multiple forms for each assessment, and in the case of computer adaptive assessments, will have very many forms (known as "test events") generated. Thus, evaluators must consider how to select the forms/events that will be subject to review in a manner that ensures the integrity and credibility of the evaluation process and results, yet that yields a feasible number of forms to review. Appendix D provides guidance on form selection.

Assessment programs with many multiple forms/events for each grade/content area may have available computer-based summaries of information suitable for informing the CCSSO Criteria evaluation. An assessment program may capture information of which forms are administered to students as part of a computer-administered program; in particular, computer-adaptive testing programs typically have this capability. Thus the methodology provides that programs may provide additional information based on computer-generated summaries/analyses of many possible or even all administered test forms/events, which may number in the several thousands. The purpose of such information would be to provide additional empirical evidence to supplement what might be learned through examination of two test forms/events. Guidance regarding this documentation is also provided in Appendix D.

CONDUCTING AN EVALUATION STUDY

The Center for Assessment has produced materials to help organizations set up and conduct an evaluation study of assessments in terms of the *CCSSO Criteria* for Test Content. These materials address many essential aspects but are not a cookbook, recognizing that an evaluation study Implementer needs to make many decisions within the particular context of the study. For example, there is no ideal number of panelists—this will depend upon the complexity and extent of the particular assessment program, the number of assessment programs reviewed, the time demands of the study, the budget, and other real constraints. Thus those contemplating conducting an evaluation study using the Test Content materials should be familiar with alignment and other content evaluation studies of assessment programs. Conversely, it is very important that any evaluation study using the Test Content methodology report on the details of who was involved, how they were qualified, the specific procedures followed, etc. to help document what was done and establish the credibility of the evaluation study's results.

ACKNOWLEDGEMENTS

The Center for Assessment's work to develop methodologies for operationalizing the *CCSSO Criteria* was funded 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 and Melinda Gates Foundation, the Lumina Foundation, the Helmsley Charitable Trust, the Charles and Lynn Schusterman Family Foundation and the William and Flora Hewlett Foundation. HQAP's resources are managed in a pooled fund by Rockefeller Philanthropy Advisors (RPA); the consulting firm Education First supports HQAP's collaborative grant making. Joanne Weiss and Judy Wurtzel worked with Education First to support this project.

CCSSO was not a formal partner in the project and had no oversight or formal involvement in the study. However, given that CCSSO developed the criteria on which the methodology is based and that state education leaders are the primary audience for this research, CCSSO provided input on the methodology to ensure its consistency with the intention of the *CCSSO Criteria* and to help ensure that the information produced in an evaluation study would address the needs of states.

The lead authors from the Center for Assessment were Brian Gong and Thanos Patelis; Karin Hess made major contributions to an early version. In the development of the Test Content, the Center for Assessment also sought input from well-known authorities on alignment methodologies, assessment content experts, and experienced staff from several assessment programs. Several persons reviewed various versions of these materials. Their frank comments helped make this document much better. They included: Tony Alpert (Smarter Balanced), Michael Briscoe (Student Achievement Partners), Brad Bolender (ACT), Bill Bryant (ACT), Lucy Calkins (Teachers College), Jayson Ceynar (ACT), Beth Cocuzza (Student Achievement Partners), Wayne Camara (ACT), Carrie Conaway (MA DESE), Magda Chia (Smarter Balanced), Sara Clough (ACT), Beth Cocuzza (Student Achievement Partners), Sheli Cole (Smarter Balanced), Colin Dingler (ACT), Enis Dogan (PARCC), Jessica Eadie (Student Achievement Partners), Haley Freeman (MA DESE), James Gambrell (ACT), Matthew Goodlaw (PARCC), Bonnie Hain (PARCC), Deborah Harris (ACT), Joan Herman (UCLA/CRESST), Joanne Jensen (WestEd), Shannon Karm (ACT), Francine Markowitz (PARCC), Donna Matovinovic (ACT), Jim Mays

(ACT), Marty McCall (Smarter Balanced), Kenneth Mullen (ACT), Scott Norton (CCSSO), Marge Petit (independent mathematics consultant), Andy Porter (University of Pennsylvania), Beth Sullivan (ACT), Doug Sovde (PARCC), Jeri Thompson (Center for Assessment), Natasha Vasavada (College Board), Norman Webb (University of Wisconsin), Andrew Wiley (Alpine Testing Solutions).

The Center for Assessment also benefited enormously from the close attention to both substance and style from those who closely reviewed the materials with an eye to practical implementation and streamlined approaches: Nancy Doorey (independent consultant on assessment), Rebecca Dvorak (HumRRO), Hilary Michaels (HumRRO), Amber Northern (Fordham), Morgan Polikoff (University of Southern California), Victoria Sears (Fordham), Sheila Shultz (HumRRO), and Carolyn Wiley (HumRRO). Any remaining deficiencies are the responsibility of the Center for Assessment authors.

MATERIALS AVAILABLE FROM THE CENTER FOR ASSESSMENT

Materials available from the Center for Assessment (www.nciea.org) in relation to the evaluation methodology for the *CCSSO Criteria* include:

- *Guide to* Evaluating Assessments Using the *CCSSO Criteria*: Focus on Test Content – Information for those wishing to understand the Test Content methodology in more depth (this document).
- Resources for those who intend to conduct an evaluation applying the Test Content methodology for the CCSSO Criteria – available by request from the Center for Assessment
- *Guide to Evaluating Assessments: Using the CCSSO Criteria: Focus on Test Characteristics* (in development)

OTHER RESOURCES

CCSSO Criteria for Procuring and Evaluating High Quality Assessments (March 2014) are available at: http://www.ccsso.org/Documents/2014/CCSSO%20Criteria%20for%20High%20Quality%20Assessments%2003242014.pdf

CCSSO Principles: States' commitment to high-quality assessments aligned to college- and career-readiness (2013) are available at: http://www.ccsso.org/Documents/2013/CCSSO%20Assessment%20Quality%20Principles%2010-1-13%20FINAL.pdf

Related documents are available from CCSSO at: http://www.ccsso.org/Resources/Programs/Assessments.html

About the Center for Assessment

The Center for Assessment is a non-partisan, non-profit 501(c)3 organization that has been providing technical assistance regarding quality assessment and accountability systems to states and other organizations since 1998. The Center for Assessment has had contracts with over 40 states and numerous school districts to provide assessment and accountability support that is technically sound, educationally beneficial, policy sensitive, and practical within the particular context of the state. The Center also is regularly invited to consult with other non-governmental and governmental organizations, including the U.S. Department of Education and the Council of Chief State School Officers. For more information, see www.nciea.org.

REFERENCES

[1] American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

APPENDIX A: SCORING SUMMARIES

List of Criteria and Sub-Criteria for English Language Arts

Sub-Criteria	Type
Criterion A.5 Providing accessibility to all 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
Criterion A.6 Ensuring transparency of test design and expectations	
A.6.1 Assessment design documents and sample test questions made publicly available	Generalizability
Criterion B.1 (Depth)	
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
Criterion B.2 (Depth)	
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
Criterion B.3 (Content)	
B.3.1 Close reading	Outcome
B.3.2 Central ideas and important particulars	Outcome
B.3.3 Questions text dependent and asses 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
Criterion B.4 (Depth)	
B.4.1 Level of cognitive demand	Outcome
B.4.2 Procedures for evaluating cognitive demand	Generalizability
Criterion B.5 (Content)	
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

Criterion B.6 (Content)	
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
Criterion B.7 (Content)	
B.7.1 Percentage 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
Criterion B.8 (Content)	
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
Criterion B.9 (Depth)	
B.9.1 Kinds of formats used on operational forms	Outcome
B.9.2 Quality of items	Outcome
B.9.3 Specifications on distribution of item types	Generalizability
B.9.4 Alignment to standards & editorial accuracy	Generalizability

A.5 Providing accessibility to *all* students, including English learners and students with disabilities (Partial)

	Type	Evidence Descriptors	Location of Evidence	Scoring Guidance	Tentative Cut-Offs
A.5.1.1	Generaliz-ability	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.).	**2 – Meets:** The assessment program has documentation regarding construct definition that is strong and comprehensive, including the following characteristics: • defines the construct to be assessed with sufficient clarity that the program and others can distinguish construct-irrelevant from construct-relevant variance; • provides a rationale for the construct definition that incorporates available research; • 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;	

				• has a process in place to improve its conception and support of validity regarding accessibility and accommodations. **1 – Partially Meets:** 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. **0 – Does Not Meet:** The assessment program's documentation manifests one or more of the following characteristics: • its definition or rationale is contrary to available research; • its definition and rationale identify the need for specific accommodations/access features but such accommodations/access features are not provided although likely practicable; • meets fewer than two of the characteristics of the 2 level. 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.	
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.).	**2 – Meets:** The assessment program has documentation that is strong and comprehensive regarding development of items with appropriate accessibility, including the following characteristics: • 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 vast majority of students; • 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) • 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; • 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.) **1 – Partially Meets:** 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. **0 – Does Not Meet:** 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. 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.	

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.).	**2 – Meets:** The assessment program has documentation that is strong and comprehensive regarding development of test forms with appropriate accessibility, including the following characteristics: • 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; • the program has procedures for assigning and delivering the appropriate accommodations/access features to individual students, including assigning special test forms; • 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; • 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. **1 – Partially Meets:** 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. **0 – Does Not Meet:** 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. 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.	

A.5.2.1	Generalizability	The assessment program offers appropriate accommodations/ access features that address the access needs of the vast 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.	**2 – Meets:** The assessment program has documentation that is strong and comprehensive regarding the accommodations/access features the program offers, including: • 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; • 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; • 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.); • 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.	

				1 – Partially Meets: 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. **0 – Does Not Meet:** 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 not adhere to its policies and procedures regarding accessibility. 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	
A.5.2.2	Outcomes	The assessment program offers appropriate accommodations/ access features that address the access needs of the vast 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.	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. 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	**2 – Meets:** 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. **1 – Partially Meets:** 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. **0 – Does Not Meet:** The Accessibility Exemplars and accompanying documentation provided by the assessment program indicates neither adequate coverage of major access/ accommodations needs nor adequate quality. 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.	

			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.		
A.5.3	Generaliz-ability	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.).	**2 – Meets:** 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. **1 – Partially Meets:** 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. **0 – Does Not Meet:** Documentation indicates the program "Does Not Meet" at least A.5.1 or A.5.2 regarding English learners. 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.	
A.5.4	Generaliz-ability	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.).	**2 – Meets:** Documentation indicates the assessment program "Meets" both (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. **1 – Partially Meets:** 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. **0 – Does Not Meet:** Documentation indicates the program "Does Not Meet" at least A.5.1 or A.5.2 regarding students with disabilities.	

				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.	

A.6 Ensuring transparency of test design and expectations					
	Type	**Evidence Descriptors**	**Location of Evidence**	**Scoring Guidance**	**Tentative Cut-Offs**
A.6.1	Generalizability	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.	Documentation submitted by assessment program.	**2 – Meets:** All of the following information is available in public documentation that is accurate and organized in a way to be accessible to stakeholders such as policy makers, state assessment program administrators, educators, and parents, and of sufficient quality to promote accurate understanding and uses of the assessments. • 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. • 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 to ensure information will remain current) across every grade level and content area. • Released items are operational items, with annotations and answer rationales provided, including scoring rubrics for constructed-response items with sample responses are provided for each level of the rubric OR the program can demonstrate that they have provided items of operational quality and associated materials that will provide the same or higher levels of information to stakeholders. • Item development specifications are provided. **1 – Partially Meets:** Some of the designated information is not available in public documentation, or information is available but of limited detail or some of the information is inaccurate or inaccessible to stakeholders. Some ways	

				information might be practically inaccessible to public stakeholders include requiring the user to compile information from across multiple documents to yield the information designated above; having information not specifically identified (e.g., having information in a table in a report that is not labeled or searchable for the designated information); not including sufficient information to interpret correctly (e.g., not clearly explaining notation or abbreviations; not clearly including significant exceptions with the information public stakeholders are likely to rely on), etc.0 – Does Not Meet: Large portions of the designated information are not available in public documentation (e.g., two or more bullets are not complete), or large portions are inaccurate and/or inaccessible to stakeholders. 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.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 Guidance	Tentative Cut-Offs
B.1.1	Outcome	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. Goals include; • In grades 3-8, approximately half of the texts are literature and half are informational. • In high school, because	Evidence: Test forms, meta-data Coding Sheets: • Is the passage informational or literary? Metrics Auto-Calculated: • Percent of passages informational. • Percent of passages literary.	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): Assign a score for grades 3-8: **2 – Meets:** Approximately half of the texts are informational. **1 – Partially Meets:** At least one-third of the texts are informational. **0 – Does Not Meet:** Less than one-third or nearly all of the texts are informational.	For grades 3 - 8: **2 – Meets:** 45-55% **1 – Partially Meets:** 33-44% or 56-84%. **0 – Does Not Meet:** 0-32% or 85-100%. For high school grades: **2 –Meets:** 60-72%. **1 – Partially Meets:** 40-59% or 73-90%. **0 – Does Not Meet:** 0-39% or 91-100%

		comprehension of complex informational texts is crucial for readiness, texts are approximately one-third literature and two-thirds informational.		Assign a score for high school: **2 –Meets:** Approximately two-thirds of the texts are informational. **1 – Partially Meets:** Less than approximately two-thirds are informational. **0 – Does Not Meet:** Less than half or nearly all of the texts are informational. 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. 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.1.2	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 • Is the passage is previously published (Y/N) • If not previously published, is the passage of publishable quality? (Y/N) Metrics Auto-Calculated: • Number/% of previously published passages • Number/% of passages of publishable quality	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): **2 –Meets:** Nearly all passages are high quality (previously published or of publishable quality). **1 – Partially Meets:** The large majority of passages (i.e. three-quarters or more) are high quality (previously published or of publishable quality).	**2 – Meets:** 90-100% **1 – Partially Meets:** 75-89% **0 – Does Not Meet:** 0-74%

				0 – Does Not Meet: 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.1.3	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: • If the passage is informational, is the structure primarily narrative or expository? (Narrative/ Expository) • 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; History/Science and Technical/ Literary Nonfiction Informational Passages)	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): **2- Meets:** 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. **1 – Partially Meets:** 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. **0 – Does Not Meet:** 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	**2 – Meets:** 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 **1 – Partially Meets:** 75-89% are expository AND/OR for grades 6-12, the informational texts address only two of the three disciplines mentioned above. **0 – Does Not Meet:** 0-74% are expository AND/OR for grades 6-12, the informational texts address only one of the three disciplines mentioned above.

			Metrics Auto-Calculated: • Number and percent of informational passages with a narrative structure • Number and percent of informational passages with an expository structure • Number and percent of history informational passages • Number and percent of science/technical informational passages • Number and percent of literary nonfiction informational passages • Number and percent of History/Literary nonfiction informational passages • Number and percent of science and technical/literary nonfiction informational passages • Number and percent of history/science and technical informational passages • Number and percent of history/science and technical/literary nonfiction informational passages	informational texts address only one of the three disciplines mentioned above. 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.1.4	Generalizability	Test blueprints and/or other specifications specify for each grade level the proportions of each text type and genre each student should be administered.	Evidence: Test blueprints and/or other documents identified by the program.	Rate the extent to which the documentation represents the distributions of the type of passages. Assign a score and provide notes under Comments: Assign a score for grades 3-8: **2 – Meets:** Specifications indicate that approximately half of the texts should be informational.	For grades 3-8: **2 –Meets:** 45-55% **1 – Partially Meets:** 33-44% or 56-84% **0 – Does Not Meet:** 0-32% or 85-100%

		The test blueprints distribution of emphasis of text types follows the CCSSO *Criteria*. Goals include: • 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. • In grades 3-8, approximately half of the texts are literature and half are informational; • In high school, texts are approximately one-third literature and two-thirds informational; • 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.		**1 – Partially Meets:** Specifications indicate that at least one-third of the texts should be informational. **0 – Does Not Meet:** Specifications indicate that less than one-third or nearly all of the texts should be informational. Assign a score for high school: **2 –Meets:** Specifications indicate that approximately two-thirds of the texts should be informational. **1 – Partially Meets:** Specifications indicate that less than approximately two-thirds should be informational. **0 – Does Not Meet:** Specifications indicate that less than half or nearly all of the texts should be informational. 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. 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.	For high school: **2 –Meets:** 60-72% **1 – Partially Meets:** 40-59% or 72-90% **0 – Does Not Meet:** 0-39% or 91-100%
B.1.5	Generalizability	As part of the construct definition, the quality of texts is defined. The program's definitions are consistent with the CCSSO *Criteria*: • Texts and other stimuli (e.g., audio,	Evidence: Test blueprints and/or other documents identified by the program.	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: **2 –Meets:** Specifications indicate that nearly all passages should be of high quality (previously	**2 – Meets:** 90-100% **1 – Partially Meets:** 75-89% **0 – Does Not Meet:** 0-74%

		visual, graphic) are previously published or of publishable quality. • They are content-rich, exhibit exceptional craft and thought, and/or provide useful information. • History/social studies and science/technical texts, specifically, reflect the quality of writing that is produced by authorities in the particular academic discipline.		published or of publishable quality). 1 – Partially Meets: 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). 0 – Does Not Meet: Specifications indicate that less than the large majority of passages should be of high quality (previously published or of publishable quality). If the writing test will not use passages, the rating will be based on reading passages only. 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.1.6	Generaliz-ability	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	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: **2- Meets:** 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.	**2 – Meets:** 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. **1 – Partially Meets:** 75-89% are expository OR for grades 6-12, the informational texts are split nearly evenly for the three disciplines mentioned above.

				1 – Partially Meets: 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. **0 – Does Not Meet:** 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. 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.	**0 – Does Not Meet:** 0-74% are expository AND for grades 6-12, the informational texts are not balanced in the three disciplines mentioned above.

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 Guidance	Tentative Cut-Offs
B.2.1	Outcome	Text complexity is quantitatively and qualitatively measured and used to place each text at the appropriate grade level. Goals include: • Texts are placed in a grade band using at least one research-based quantitative measure;	Evidence: Test forms, meta-data Coding Sheet • Is there evidence of both quantitative and qualitative analysis? (Y/N) • Is the passage placed in appropriate grade band based on quantitative data? (Y/N or N/A) • Is the passage placed in appropriate grade level based on qualitative analysis? (Y/N)	Determine the percentage of passages placed at a grade band that is justified by quantitative data and a grade level justified by qualitative measures. Assign a score and provide notes under Comments (for each form): **2 – Meets:** All or nearly all passages have been placed at a grade band and grade level justified by complexity data. **1 – Partially Meets:** 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.	**2 – Meets:** 90-100% **1 – Partially Meets:** 75-89% **0 – Does Not Meet:** 0-74%

		• Texts are placed at a grade level using a qualitative analysis measure, reflecting the expert judgment of educators; and • Most of the texts are placed within the grade band indicated by the quantitative analysis, with exceptions usually found in high school literary texts See Common Core State Standards Appendix A regarding text complexity.	Metrics Auto-Calculated: • Number and percent of texts placed in correct grade band based on quantitative data • Number and percent of texts placed in correct grade level based on qualitative data • Number and percent of texts placed in correct grade band based on quantitative data AND in correct grade level based on qualitative analysis	**0 – Does Not Meet:** Less than a large majority of passages have been placed at a grade band justified by complexity data "Complexity data" refers to results from both quantitative and qualitative measures. 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.2.2	Generaliz-ability	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. Goals include: • Texts are placed in a grade **band** using at least one research-based quantitative measure; • Texts are placed at a grade **level** using a qualitative analysis measure, reflecting the expert judgment of educators; and • Most of the texts are placed within the grade **band** indicated by the quantitative analysis, with exceptions usually found in high school literary texts.	Evidence: Test blueprints and/or other documents identified by the program.	Evaluate whether the documentation indicates the percentage of passages placed at a grade band that is justified by quantitative data and a grade level justified by qualitative measures. Assign a rating and provide notes under Comments: **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.* **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.*	**2 – Meets:** 90-100% **1 – Partially Meets:** 75-89% **0 – Does Not Meet:** 0-74%

				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.* *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). 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.2.3	Generaliz-ability	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.	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: **2 –Meets:** Documentation outlines that text complexity increases by grade level across all years of the assessment program, meeting CCR levels by end of high school. **1 – Partially Meets:** Documentation outlines that text complexity increases by grade band across all years of the assessment program, meeting CCR levels by end of high school.	**2 – Meets:** details progression by grade level **1 – Partially Meets:** details progression by grade band only **0 – Does Not Meet:** does not include details about increasing text complexity

				0 – Does Not Meet: 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. 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.3 Requiring students to read closely and use evidence from texts: 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-Offs
B.3.1	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: • Assigned CCSS alignment (and secondary alignment(s), if any) Point value of item Coding Sheets: • Is the item aligned to the specifics of the standard? (Y/N) • Does item require close reading and analysis? (Y/N) • Does item focus on central ideas and important particulars? (Y/N) • Does the item require direct use of textual evidence? (Y/N) Metrics Auto-Calculated: • Total reading items	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): **2 – Meets:** Nearly all items require close reading and analysis of text. **1 – Partially Meets:** The large majority of items (i.e., three-quarters or more) require close reading and analysis of text. **0 – Does Not Meet:** 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.	**2 – Meets:** 90-100% **1 – Partially Meets:** 75-89% **0 – Does Not Meet:** 0-74%

B.3.2	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.	• Total reading score points • Number and percent of items aligned to the specifics of the standard • Number and percent of the items requiring close reading. • Number and percent of the items focusing on central ideas • Number and percent of the items requiring direct textual evidence • Number and percent of the reading score points requiring direct textual evidence	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): **2 – Meets:** Nearly all the items focus on central ideas and important particulars **1 – Partially Meets:** The large majority of items (i.e., three-quarters or more) focus on central ideas and important particulars. **0 – Does Not Meet:** 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.	**2 – Meets:** 90-100% **1 – Partially Meets:** 75-89% **0 – Does Not Meet:** 0-74%
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).		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): **2 – Meets:** Nearly all items are aligned to the specifics of the standards. **1 – Partially Meets:** The large majority of items (i.e., three-quarters or more) are aligned to the specifics of the standards. **0 – Does Not Meet:** Less than the large majority of the items are aligned to the specifics of the standards. 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	**2 – Meets:** 90-100% **1 – Partially Meets:** 75-89% **0 – Does Not Meet:** 0-74%

				Standards should be assigned a "0" and rated as Does Not Meet to this metric. 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.3.4	Outcome	Many reading questions require students to directly provide textual evidence in support of their responses. Goals include: • 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).		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): **2 – Meets:** More than half of the reading score points are based on items requiring direct use of textual evidence. **1 – Partially Meets:** Nearly half of the score points are based on items requiring direct use of textual evidence. **0 – Does Not Meet:** Less than one-third of the score points are based on items requiring direct use of textual evidence. 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.	**2 – Meets:** 51-100% **1 – Partially Meets:** 33-50% **0 – Does Not Meet:** 0-32%
B.3.5	Generalizability	Item specifications require all reading questions to be text-dependent. They require that reading questions: • Arise from and require close reading and analysis of text;	Evidence: Test blueprints and/or other documents identified by the program.	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: **2 – Meets:** Documentation outlines expectations for items	**2 – Meets:** All three **1 – Partially Meets:** Two of three **0 – Does Not Meet:** One of three

		• Focus on the central ideas and important particulars of the text, rather than on superficial or peripheral concepts; 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.		to require close reading AND to focus on central ideas and important particulars, AND align to the specifics of the standards. **1 – Partially Meets:** Documentation outlines expectations for only two of the three emphases mentioned above. **0 – Does Not Meet:** Documentation outlines expectations for one or none of the emphases mentioned above. 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.3.6	Generalizability	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).	Evidence: Test blueprints and/or other documents identified by the program.	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: **2 – Meets:** Documentation indicates that more than half of the reading score points should be based on items requiring direct use of textual evidence. **1 – Partially Meets:** Documentation indicates that half or less of score points should be based on items requiring direct use of textual evidence. **0 – Does Not Meet:** Documentation indicates that less than one-third of the score points should be based on items requiring direct use of textual evidence.	**2 – Meets:** 51-100% **1 – Partially Meets:** 33-50% **0 – Does Not Meet:** 0-32%

				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.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
B.4.1	Outcome	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: • Complexity of the text on which an item is based; • Range of textual evidence an item requires (how many parts of text[s] students must locate and use to response to the item correctly); • Level of inference required; and • Mode of student response (e.g., selected-response, constructed-response).	Evidence: Test forms Specific metadata from assessment program: • Point value of item • Assigned CCSS alignment (multiple standards shown, if applicable) • If program uses Webb, assigned item DoK • If program does not use Webb, assigned item cognitive demand level Coding Sheets: • By Standard: primary DoK, secondary DoK, tertiary DoK, quaternary DoK. • By item: Indicate DoK Metrics Auto-Calculated: For each test form: • Number and percent of standards at each of the DoK levels • DoK Index = comparing the percentage of score points for items at each DoK level with the percentage of standards at that DoK level, identifying	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). **2 –Meets:** 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. **1 – Partially Meets:** 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. **0 – Does Not Meet:** 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	**2 – Meets:** • The DoK Index is at least 80% AND the percentage of score points associated with DoK3+ items is no more than 10% less than the percentage of standards that are DoK3+. **1 – Partially Meets:** • The DoK Index is at least 60% AND the percent of DoK1 score points is no more than 20% higher than the percentage of standards that are DoK1. **0 – Does Not Meet:** • The DoK Index is less than 60% OR the percent of DoK1 score points is more than 20% greater than the percentage of standards that are DoK1.

			whichever is less, and summing the percentages of the minima • DoK Index averaged across both test forms.	rationale for insufficient information determination. For example, one or more pieces of evidence listed in the "Location of Evidence" column were not available.	
B.4.2	Generaliz-ability	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: • Complexity of the text on which an item is based; • Range of textual evidence an item requires (how many parts of text[s] students must locate and use to response to the item correctly); • Level of inference required; and • Mode of student response (e.g., selected-response, constructed-response).	Evidence: Test blueprints and/or other documents identified by the program.	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. **2–Meets:** 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. **1 – Partially Meets:** 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. However, one or more of these pieces of information is inadequately described or justified. 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. **0 – Does Not Meet:** Documentation does not indicate a definition of cognitive demand, a way of operationalizing cognitive demand at the item level,	**2 – Meets:** • If the program uses Webb, the DoK Index is at least 80% AND • the percentage of score points associated with DoK3+ items is no more than 10% less than the percentage of standards that are DoK3+. • 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. **1 – Partially Meets:** • If the program uses Webb, the DoK Index is at least 60% AND • the percent of DoK1 score points is no more than 20% higher than the percentage of standards that are DoK1.

				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. 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.	• 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. • However, one or more of these pieces of information is inadequately described or justified. **0 – Does Not Meet:** • If the program uses Webb, the DoK Index is less than 60% OR • the percent of DoK1 score points is more than 20% greater than the percentage of standards that are DoK1. • 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.

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 Guidance	Tentative Cut-Offs
B.5.1	Outcome	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. Goals include: • 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.	Evidence: Test forms, meta-data. Specific metadata from assessment program: • Assigned CCSS alignment(and secondary alignment(s), if any) • Point value of item • Chart indicating types of writing assessed at each grade level in the grade band Coding Sheet: • What type of writing is called for? (Expository; Persuasive/ argumentative; Narrative; Blended) Coding Sheet Auto calculation: • Total number of writing items • Number and percent of CRs requiring expository writing • Number and percent of CRs requiring persuasive/ argumentative writing • Number and percent of CRs requiring narrative writing • Number and percent of CRs requiring blended writing	Determine the percentages of prompts requiring writing to sources. Assign a score and provide notes under Comments: **For grades 3 -8 and for high school programs that test narrative writing:** **2 – Meets:** All three writing types are approximately equally represented across all forms in the grade band, allowing blended types to contribute to the distribution **1 – Partially Meets:** Two of the three writing types are represented across all forms in the grade band, allowing blended types to contribute to the distribution. **0 – Does Not Meet:** One of the three writing types is represented across all forms in the grade band. NOTE: If the high school assessments do not include narrative writing, the assessment can still be rated as Meets. **For high school programs that do NOT include narrative writing:** **2 – Meets:** Expository and argument writing types are approximately equally represented across all forms in the grade band, allowing blended types to contribute to the distribution	For grades 3 -8 and for high school programs that test narrative writing: **2 – Meets:** 28-38% of each representing exposition, argument, and narrative **1 – Partially Meets:** Two of the three writing types are present and one type is 0%-27% **0 – Does Not Meet:** One type is 100% **For high school programs that do NOT include narrative writing:** **2 – Meets:** 40-60% each for expository and argument types. **1 – Partially Meets:** Both expository and argument types are represented, but one writing type accounts for more than 60% of the balance of these two types. **0 – Does Not Meet:** Either expository or argument is not represented, or neither is represented.

				1 – Partially Meets: Both writing types are represented but one much more heavily than the other **0 – Does Not Meet:** Only one or no writing type (expository OR argument) is represented. 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.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.	Evidence: Test forms, meta-data Specific metadata from assessment program: • Assigned CCSS alignment (and secondary alignment(s), if any) • Point value of item. Coding Sheet: • Is the writing task text-based? (Y/N) Coding Sheet Auto calculation: •Total number of writing items • Number and percent of text-based writing tasks	Determine the percentages of prompts requiring writing to sources. Assign a score and provide notes under Comments (for each form): **2 – Meets:** All writing prompts require writing to sources (are text-based). **1 – Partially Meets:** The large majority (i.e., three-quarters or more) of writing prompts require writing to sources (are text-based). **0 – Does Not Meet:** Fewer than the large majority of writing prompts require writing to sources (are text-based) OR the program does not include writing prompts. 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.	**2 – Meets:** 90-100% **1 – Partially Meets:** 75-89% **0 – Does Not Meet:** 0-74%
B.5.3	Generalizability	Test blueprints and/or other specifications specify the distribution of the various writing tasks/types as standards require, and at higher	Evidence: Test blueprints and/or other documents identified by the program.	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	**For grades 3 -8 and for high school programs that test narrative writing:** **2 – Meets:** 28-38% of each representing

		grade levels the balance shifts toward more exposition and argument. Goals include: • 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.		**For grades 3 -8 and for high school programs that test narrative writing:** **2 –Meets:** Documentation indicates that all three writing types are approximately equally represented in the grade band, allowing blended types to contribute to the distribution. **1 – Partially Meets:** Documentation indicates that two of the three writing types are represented in the grade band, allowing blended types to contribute to the distribution **0 – Does Not Meet:** Documentation indicates that one of the three writing types is represented in the grade band. NOTE: If the high school assessments do not include narrative writing, the assessment can still be rated as aligned. **For high school programs that do NOT include narrative writing:** **2 – Meets:** 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 **1 – Partially Meets:** 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. **0 – Does Not Meet:** Documentation indicates that only one writing type (expository OR argument) should be represented. Insufficient information box checked if there is insufficient information to score. Comments must be added to explain	exposition, argument, and narrative **1 – Partially Meets:** Two of the three writing types are present and one type is 0%-27% **0 – Does Not Meet:** One type is 100% **For high school programs that do NOT include narrative writing:** **2 – Meets:** 40-60% each for expository and argument types. **1 – Partially Meets:** Both expository and argument types are represented, but one writing type accounts for more than 60% of the balance of these two types. **0 – Does Not Meet:** Either expository or argument is not represented, or neither is represented.

				rationale for insufficient information determination. For example, one or more pieces of evidence listed in the “Location of Evidence” column were not available.	
B.5.4	Generaliz-ability	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.	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. **2 –Meets:** Documentation indicates that all writing prompts require writing to sources (are text-based). **1 – Partially Meets:** Documentation indicates that the large majority (i.e., three-quarters or more) of writing prompts require writing to sources (are text-based). **0 – Does Not Meet:** 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. 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.	**2 – Meets:** 90-100% **1 – Partially Meets:** 75-89% **0 – Does Not Meet:** 0-74%

B.6 Emphasizing vocabulary and language skills: The assessments require students to demonstrate proficiency in the use of language, including vocabulary and conventions.

	Type	Evidence Descriptors	Location of Evidence	Scoring Guidance	Tentative Cut-Offs
B.6.1	Outcome	Vocabulary items reflect requirements for college and career readiness, including focusing on general	Evidence: Test forms, meta-data Specific metadata from assessment program: • Point value for item	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	**2 – Meets:** 75-100% Tier 2; 51% -100% Central **1 – Partially Meets:** 50-75% Tier 2; 33-50% Central

		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.	• Primary CCSS alignment • Any Secondary CCSS alignment Coding Sheet: • Does the item test a Tier 2 Academic word or phrase? (Y/N) • Does the item test a word central to the understanding of the text? (Y/N) • Does the tested word require use of context? (Y/N) Coding Sheet Auto calculation: • Total vocabulary items • Total vocabulary points • Number and percent of items testing Tier 2 words or phrases • Number and percent of vocabulary items testing words/phrases central to the text • Number and percent of vocabulary items requiring context Number and percent of vocabulary items testing Tier 2 words or phrases AND requiring context	notes under Comments (for each form): 2 – Meets: 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. 1 – Partially Meets: 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. 0 – Does Not Meet: 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. 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. 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.	**0 – Does Not Meet:** 0-49% Tier 2; 0-32% Central
B.6.2	Outcome	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 mirroring real-world	Evidence: Test forms, meta-data, and writing rubric Specific metadata from assessment program: • Assigned CCSS alignment (and secondary alignment(s), if any) • Score points for each item	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): **2 – Meets:** 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-	**2 – Meets:** 75-100% **1 – Partially Meets:** 50-74% **0 – Does Not Meet:** 0-49%

		activities (e.g., actual editing or revision, actual writing); and focusing on common student errors and those conventions most important for readiness.	Coding Sheet: • Does item mirror real-world activities? (Y/N) • Does item test conventions most important for readiness (see CCSS Language Skills Progression Chart)? (Y/N) • Does the item focus on common student errors? (Y/N) Coding Sheet Auto calculation: • Total language items • Total language score points • Number and percent of reading items that mirror real-world activities • Number and percent of items that test conventions most important for readiness • Number and percent of items that focus on common student errors	world activities, focus on common errors, and emphasize the conventions most important for readiness. **1 – Partially Meets:** 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. **0 – Does Not Meet:** 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. 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.6.3	Outcome	Assessments place sufficient emphasis on vocabulary (i.e., a significant percentage of the score points is devoted to these skills)	Evidence: Test forms, metadata Specific metadata from assessment program: • Assigned CCSS alignment (and secondary alignment(s), if any) • Score points for each item Coding Sheet Auto calculation: • Number and percent of score points devoted to assessing vocabulary	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): **2 – Meets:** Vocabulary is reported as a subscore OR at least 13% of score points are devoted to assessing vocabulary **1 – Partially Meets:** At least 10%of score points are devoted to assessing vocabulary **0 – Does Not Meet:** Less than 10% of points are devoted to assessing vocabulary	**2 – Meets:** Vocabulary is reported as a subscore OR > 13% of score points **1 – Partially Meets:** 10 -12% of score points **0 – Does Not Meet:** 0 to 9% of score points

				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.6.4	Outcome	Assessments place sufficient emphasis on language skills (i.e., a significant percentage of the score points is devoted to these skills)	Evidence: Test forms, meta-data, and writing rubric Specific metadata from assessment program: • Assigned CCSS alignment (and secondary alignment(s), if any) • Score points for each item Coding Sheet Auto calculation: • Number and percentage of score points devoted to assessing language.	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): **2 – Meets:** Language skills are reported as a subscore OR 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). **1 – Partially Meets:** At least 10% of score points are devoted to assessing language skills **0 – Does Not Meet:** Less than 10% of points are 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.	**2 – Meets:** Language skills are reported as a subscore OR >13% of score points **1 – Partially Meets:** 10-12% of score points **0 – Does Not Meet:** 0 to 9% of score points

B.6.5	Generalizability	Item specifications require that 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.	Evidence: Test blueprints and/or other documents identified by the program.	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: **2 – Meets:** 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. **1 – Partially Meets:** 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. **0 – Does Not Meet:** 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. 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. 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.	**2 – Meets:** 75-100% tier 2 and require use of context; and 51% -100% Central **1 – Partially Meets:** 50-74% tier 2 and require use of context; and/or 33-50% Central **0 – Does Not Meet:** 0-49% tier 2 and require use of context; 0-32% Central
B.6.6	Generalizability	Item specifications require that language is assessed within writing assessments as part of the scoring rubric, or it is assessed with test	Evidence: Test blueprints and/or other documents identified by the program.	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	**2 – Meets:** 75-100% **1 – Partially Meets:** 50-74% **0 – Does Not Meet:** 0-49%

		items that specifically address language skills. 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.		notes under Comments: **2 – Meets:** 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. **1 – Partially Meets:** 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. **0 – Does Not Meet:** 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. 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.6.7	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.	Determine the percentage of score points associated with vocabulary to support sufficient emphasis and provide notes under Comments: **2 – Meets:** Documentation indicates that vocabulary is reported as a subscore OR at least 13% of score points should be devoted to assessing vocabulary.	**2 – Meets:** Vocabulary is reported as a subscore or > 13% of score points **1 – Partially Meets:** 10=12% of score points **0 – Does Not Meet:** 0 to 9% of score points

				1 – Partially Meets: Documentation indicates that at least 10% of score points should be devoted to assessing vocabulary. **0 – Does Not Meet:** 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.6.8	Generalizability	• Assessments place sufficient emphasis on vocabulary and language skills (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.	Determine the percentage of score points devoted to language skills and provide notes under Comments: **2 – Meets:** 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). **1 – Partially Meets:** Documentation indicates that at least 10% of score points should be devoted to assessing language skills. **0 – Does Not Meet:** 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.	**2 – Meets:** Language skills are reported as a subscore OR > 13% of score points **1 – Partially Meets:** 10-12% of score points **0 – Does Not Meet:** 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 Guidance	Tentative Cut-Offs
B.7.1	Outcome	Test items assessing research and inquiry mirror real world activities and require students to analyze, synthesize, organize, and use information from sources. Goals include: • Research tasks require writing to sources, including analyzing, selecting, and organizing evidence from more than one source, and often from sources in diverse formats	Evidence: Test forms, meta-data Specific metadata from assessment program: • Point value • Grade level • Primary assigned CCSS alignment • Any secondary CCSS alignment Coding Sheet: • Does item require analysis, synthesis, and/or organization of information (mirroring real-world activities)? (Y/N) Coding Sheet Auto calculation: • Total research items • Total research score points • Number and percent of items mirroring real world activities • Number and percent of items devoted to research • Number and percent of points devoted to research	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): **2 – Meets:** The large majority (i.e., three-quarters or more) of the research items require analysis, synthesis, and/or organization of information. **1 – Partially Meets:** More than half of the research items require analysis, synthesis, and/or organization of information. **0 – Does Not Meet:** Half or less than half of research items require analysis, synthesis, and/or organization of information 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. 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.	**2 – Meets:** 75-100% **1 – Partially Meets:** 51-74% **0 – Does Not Meet:** 0-50%
B.7.2	Generalizability	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: • When assessment constraints permit,	Evidence: Test blueprints and/or other documents identified by the program.	Determine the percentage of score points assessing real or simulated research tasks. Assign a score and provide notes under Comments: **2 – Meets:** Program reports a research score or otherwise demonstrates that research is significant. **1 – Partially Meets:** Program includes research items should be assessed but these are not	**2 – Meets:** Program reports a research score or otherwise demonstrates that research is significant. **1 – Partially Meets:** Program includes research items should be assessed but these are not reported or indicates research is not significant.

		real or simulated research tasks comprise a significant percentage of score points when all forms of the reading and writing test are considered together.		reported or program does not indicate research is significant. **0 – Does Not Meet:** No research items are specified to be included. Note: A research item, at a minimum, includes paired nonfiction passages with a writing task. 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.	**0 – Does Not Meet:** No research items are specified to be included.
B.7.3	Generaliz-ability	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: • Research tasks require writing to sources, including analyzing, selecting, and organizing evidence from more than one source, and often from sources in diverse formats.	Evidence: Test blueprints and/or other documents identified by the program.	Determine the percentage of test items assessing research and inquiry mirroring real world activities. Assign a score and provide notes under Comments: **2 – Meets:** Documentation indicates that the large majority (i.e., three-quarters or more) of the research items require analysis, synthesis, and/or organization of information. **1 – Partially Meets:** Documentation indicates that more than half of the research items require analysis, synthesis, and/or organization of information **0 – Does Not Meet:** Documentation indicates that half or less than half of research items require analysis, synthesis, and/or organization of information. NOTES: If there is no research component, rate this evidence descriptor as 0. If the assessment offers paired nonfictional passages with a writing task, count that section of the test as research.	**2 – Meets:** 75-100% **1 – Partially Meets:** 51-74% **0 – Does Not Meet:** 0-50%

				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.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 Guidance**	**Tentative Cut-Offs**
B.8.1	Outcome	Over time, and as assessment advances allow, the listening skills required for college and career readiness are assessed. Test items assessing listening: • 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 • Permit the evaluation of active listening skills (e.g., taking notes on main ideas, elaborating on remarks of others).	Evidence: Test forms, meta-data Specific metadata from assessment program: • Assigned CCSS alignment (and any secondary alignment(s), if any) Coding Sheet: • Does listening stimulus meet expectations for quality as outlined in B.1? (Y/N) "B.8.1 • Does the listening stimulus meet the expectations for complexity outlined in B2? (Y/N) • Does listening item require active listening? (Y/N) Coding Sheet Auto calculation: • Total listening items • Number and percent of listening items with stimuli that meet B.1 & B.2 expectations for complexity and quality • Number and percent of listening items that require active listening	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). **2 – Meets:** 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. **1 – Partially Meets:** Many (i.e., at least half) of listening items meet the requirements outlined in B.1 and B.2 AND evaluate active listening skills. **0 – Does Not Meet:** 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. 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.	**2 – Meets:** 75-100% **1 – Partially Meets:** 50-74% **0 – Does Not Meet:** 0-49%

			• Number and percent of listening items that require active listening AND with stimuli that meet B.1 & B.2 expectations for complexity and quality		
B.8.2	Outcome	Over time, and as assessment advances allow, the speaking skills required for college and career readiness are assessed. Test items assessing speaking: • Assess students' ability to express well-supported ideas clearly and to probe others' ideas; and • Include items that measure students' ability to marshal evidence from research and orally present findings in a performance task.	Evidence: Test forms, meta-data Specific metadata from assessment program: • Assigned CCSS alignment Coding Sheet: • Does the item assess student's ability to express well supported ideas clearly and to probe other's ideas? (Y/N) • Does the item measure students' ability to marshal evidence from research? (Y/N) • Does the item measure students' ability to orally present findings? (Y/N) Coding Sheet Auto calculation: • Number and percent of speaking items assessing students' ability to express well supported ideas and probe others ideas • Number and percent of speaking items that measure students ability to marshal evidence from research. • Number and percent of speaking items that measure students' ability to orally present findings.	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). **2 – Meets:** 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. **1 – Partially Meets:** 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. **0 – Does Not Meet:** 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. 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.	**2 – Meets:** 75-100% **1 – Partially Meets:** 50-74% **0 – Does Not Meet:** 0-49%

			• 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.		
B.8.3	Generaliz-ability	Item specifications and other ancillary documents specify that test items assessing listening reflect current assessment capabilities and constraints. Test items assessing listening: • 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 • Permit the evaluation of active listening skills (e.g., taking notes on main ideas, elaborating on remarks of others).	Evidence: Test blueprints and/or other specification documents.	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: **2 – Meets:** 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. **1 – Partially Meets:** 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. **0 – Does Not Meet:** 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. 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.	**2 – Meets:** 75-100% **1 – Partially Meets:** 50-74% **0 – Does Not Meet:** 0-49%

B.8.4	Generaliz-ability	Item specifications and other ancillary documents specify that test items assessing speaking reflect current assessment capabilities and constraints. Test items assessing speaking: • Assess students' ability to express well-supported ideas clearly and to probe others' ideas; and • Include items that measure students' ability to marshal evidence from research and orally present findings in a performance task.	Evidence: Test blueprints and/or other specification documents.	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: **2 – Meets:** 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. **1 – Partially Meets:** 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. **0 – Does Not Meet:** 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. 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.	**2 – Meets:** 75-100% **1 – Partially Meets:** 50-74% **0 – Does Not Meet:** 0-49%

B.9 Ensuring high-quality items and a variety of item types: 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.	Evidence: Test forms, meta-data Specific metadata from assessment program: • Item type Coding Sheet: • Are there 2 or more item types? (Y/N) Does at least one of the item types require students to generate, rather than select, a response? (Y/N) Coding Sheet auto calculation: • Number and percent of multiple choice items • Number and percent of multi-select items • Number and percent of evidence-based selected response items • Number and percent of technology enhanced items (does not require student to generate a response) • Number and percent of constructed/student generated responses • Number and percent of items with other item type • Number and percent of high quality items	Determine the kinds of item formats used on the operational forms. Assign a score and provide notes under Comments (for each form): **2 – Meets:** At least two item formats are used, including one that requires students to generate, rather than select, a response (i.e., CR, extended writing). **1 – Partially Meets:** At least two formats (but not including CR) are used, including technology-based formats and/ or two-part selected response formats. **0 – Does Not Meet:** Only a traditional multiple choice format is used. 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.	**2 – Meets:** At least two item formats are used, including one that requires students to generate, rather than select, a response (i.e., CR, extended writing). **1 – Partially Meets:** At least two formats (but not including CR) are used, including technology-based formats and/or two-part selected response formats. **0 – Does Not Meet:** Only a traditional multiple choice format is used.
B.9.2	Outcome	Operational items are reviewed to verify claims of quality, including ensuring the technical quality, alignment to standards, and	Evidence: Test forms, meta-data Specific metadata from assessment program: • Point value of item	Using the provided documentation, determine that there are high-quality items. Assign a score and provide notes under Comments (for each form):	**2 – Meets:** 95-100% for editorial and technical; 90% for alignment to standards **1 – Partially Meets:** 90-94% for

		editorial accuracy of the items.	• Assigned CCSS alignment Coding Sheets: • Do you agree with the assigned CCSS Alignment? (Y/N) • Is there a quality issue with this item? (Y/N) • If so, what is the issue? (Select all that apply) - Item may not yield valid evidence of targeted skill - Item has issues with readability - Item incorrectly keyed - Item has unintended correct answer - Content is inaccurate - Item has issues with editorial accuracy Metrics auto-calculated: • % of high-quality items • % of agreement with given alignment	**2 –Meets:** All or nearly all operational items reviewed reflect technical quality, alignment to standards, and editorial accuracy. **1 – Partially Meets:** A few operational items reviewed have issues with technical quality, alignment to standards, and/or editorial accuracy. **0 – Does Not Meet:** 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. Note: Reviewers may enter comments about the quality of specific items in the Item Worksheet. 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.	editorial and technical; 80% for alignment to standards **0 – Does Not Meet:** 0-89% for editorial and technical; 0-79% for alignment to standards
B.9.3	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.	Assign a score representing the specification for ensuring high-quality items and a variety of item types; provide notes under Comments: **2 – Meets:** 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). **1 – Partially Meets:** 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.	**2 – Meets:** 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). **1 – Partially Meets:** Specifications indicate that at least two formats (but not including CR) should be used, including technology-based formats and/or

				0 – Does Not Meet: Documentation indicates that only a single format should be used, including traditional multiple-choice format. 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.	two-part selected response formats. **0 – Does Not Meet:** Specifications indicate that only a single format should be used, including traditional multiple-choice format.
B.9.4	Generalizability	To support claims of quality, the following are provided in documentation: • Rationales for the use of the specific item types; • Specifications showing the proportion of item types on a form; • 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; A description of the process used for ensuring the technical quality, alignment to standards, and editorial accuracy of the items.	Evidence: Test blueprints, administration and scoring manuals, QC procedure documents, and/or other documents provided by the program.	Assign a score and provide notes under Comments: **2 –Meets:** Documentation supports claims of the technical quality, alignment to standards, and editorial accuracy. **1 – Partially Meets:** Documentation partially supports claims of the technical quality, alignment to standards, and/or editorial accuracy. **0 – Does Not Meet:** Documentation does not support claims of the technical quality, alignment to standards, and/or editorial accuracy. 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.	**2 –Meets:** Documentation supports claims of the technical quality, alignment to standards, and editorial accuracy. **1 – Partially Meets:** Documentation partially supports claims of the technical quality, alignment to standards, and/or editorial accuracy. **0 – Does Not Meet:** Documentation does not support claims of the technical quality, alignment to standards, and/or editorial accuracy.

Scoring Summary for English Language Arts

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
			❑: Missing			
		Comment:				
		A.5.1.2				
			❑: Missing			
		Comment:				
		A.5.1.3				
			❑: Missing			
		Comment:				
A.5.2	Offering appropriate accommodations/access features	A.5.2.1				
			❑: Missing			
		Comment:				
A.5.3	English learners	A.5.3				
			❑: Missing			
		Comment:				
A.5.4	Students with disabilities	A.5.4				
			❑: 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		
			❑: Missing			
		Comment:				

Criterion		Sub-Criterion	Score	Automatic Criterion-Level Raw Score	Group Criterion Score Rules
A.6.1	Information available to the public	A.6.1	(0/1/2 Score)		E G L W
			❑: Missing		
		Comment:			

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			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			Add (0/1/2) ratings from each form and each outcome sub-criterion. Range: 0 to 12	E G L W
			❑: Missing	❑: Missing			❑: Missing	❑: Missing		
		B.1.2								
			❑: Missing	❑: Missing			❑: Missing	❑: Missing		
		B.1.3								
			❑: Missing	❑: Missing			❑: Missing	❑: 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			
					❑: Missing					
		B.1.5			(0/1/2) Rating					
					❑: Missing					
		B.1.6			(0/1/2) Rating					
					❑: Missing					
		Comments:								
B.2	Focusing on complexity of texts	B.2.1			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								
		B.2.2					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.2.3								
		Comments								

B.3	**Requiring students to read closely and use evidence from texts**	**B.3.1**			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
			❑: Missing	❑: Missing			❑: Missing	❑: Missing		
		B.3.2								
			❑: Missing	❑: Missing			❑: Missing	❑: Missing		
		B.3.3								
			❑: Missing	❑: Missing			❑: Missing	❑: Missing		
		B.3.4								
			❑: Missing	❑: Missing			❑: Missing	❑: Missing		
		Comments								
		B.3.5			(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			
					❑: Missing					
		B.3.6			(0/1/2) Rating					
					❑: Missing					
		Comments								
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
			❑: Missing	❑: Missing			❑: Missing	❑: 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			
					❑: 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
			❑: Missing	❑: Missing			❑: Missing	❑: Missing		
		B.5.2								
			❑: Missing	❑: Missing			❑: Missing	❑: 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			
					❑: Missing					
		B.5.4			(0/1/2) Rating					
					❑: Missing					
		Comments								
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
			❑: Missing	❑: Missing			❑: Missing	❑: Missing		
		B.6.2								
			❑: Missing	❑: Missing			❑: Missing	❑: Missing		
		B.6.3								
			❑: Missing	❑: Missing			❑: Missing	❑: Missing		
		B.6.4								
			❑: Missing	❑: Missing			❑: Missing	❑: 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			
					❑: Missing					
		B.6.6			(0/1/2) Rating					
					❑: Missing					
		B.6.7			(0/1/2) Rating					
					❑: Missing					
		B.6.8			(0/1/2) Rating					
					❑: Missing					
		Comments								

B.7	**Assessing research and inquiry**	**B.7.1**			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								
		B.7.2			(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					
		B.7.3			(0/1/2) Rating					
					❑: Missing					
		Comments								
B.8	**Assessing speaking and listening**	**B.8.1**			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			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								
		B.8.2								
			❑: Missing	❑: Missing			❑: Missing	❑: Missing		
		Comments								
		B.8.3			(0/1/2) Rating					
					❑: Missing					
		Comments								
		B.8.4								
					❑: Missing					
		Comments								
B.9	**Ensuring high-quality items and a variety of item types**	**B.9.1**			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
			❑: Missing	❑: Missing			❑: Missing	❑: Missing		
		B.9.2								
			❑: Missing	❑: Missing			❑: Missing	❑: Missing		
		Comments								
		B.9.3			(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			
					❑: Missing					
		Comments								

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, 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 B1 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.

List of Criteria and Sub-Criteria for Mathematics

Criteria & Sub-Criteria	Type
Criterion A.5 Providing accessibility to all 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
Criterion A.6 Ensuring transparency of test design and expectations	
A.6.1 Assessment design documents and sample test questions made publicly available	Generalizability
Assesses the content most needed for College and Career Readiness (Cluster)	
Criterion C.1 Focusing strongly on the content most needed for success in later mathematics	
C.1.1 Most important content assessed	Outcome
C.1.2 Assessment design reflect important content	Generalizability
C.1.3 The assessment design reflects the standards and reflects a coherent progression of mathematics content from grade to grade and course to course.	Generalizability
Criterion C.2 Assessing a balance of concepts, procedures, and applications	
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
Assesses the depth that reflect the demands of College and Career Readiness (Cluster)	
Criterion C.3 Connecting practice to content	
C.3.1 Meaningful connections between practices and content	Outcome
C.3.2 Specifications & explanation of assessing math practices with content	Generalizability
C.3.3 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).	Generalizability
Criterion C.4 Requiring a range of cognitive demand	
C.4.1 Cognitive Demand	Outcome
C.4.2. Specification of Cognitive Demand	Generalizability
Criterion C.5 Ensuring high-quality items and a variety of item types	
C.5.1 Distribution of item types	Outcome
C.5.2 Degree of high-quality items	Outcome
C.5.3 Specification of item types and quality	Generalizability
C.5.4 Specification of distribution of item types	Generalizability

A.5 Providing accessibility to *all* students, including English learners and students with disabilities (Partial)					
	Type	**Evidence Descriptors**	**Location of Evidence**	**Scoring Guidance**	**Tentative Cut-Offs**
A.5.1.1	Generaliz-ability	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.).	**2 – Meets:** The assessment program has documentation regarding construct definition that is strong and comprehensive, including the following characteristics: • defines the construct to be assessed with sufficient clarity that the program and others can distinguish construct-irrelevant from construct-relevant variance; • provides a rationale for the construct definition that incorporates available research; • 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; • has a process in place to improve its conception and support of validity regarding accessibility and accommodations. **1 – Partially Meets:** 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. **0 – Does Not Meet:** The assessment program's documentation manifests one or more of the following characteristics: • its definition or rationale is contrary to available research; • its definition and rationale identify the need for specific accommodations/ access features but such accommodations/access features are not provided although likely practicable; • meets fewer than two of the characteristics of the 2 level. 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.	

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.).	**2 – Meets:** The assessment program has documentation that is strong and comprehensive regarding development of items with appropriate accessibility, including the following characteristics: • 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 vast majority of students; • 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) • 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; • 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.)	

				1 – Partially Meets: 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. **0 – Does Not Meet:** 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. 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.	

A.5.1.3	Generaliz-ability	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.).	**2 – Meets:** The assessment program has documentation that is strong and comprehensive regarding development of test forms with appropriate accessibility, including the following characteristics: • 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; • the program has procedures for assigning and delivering the appropriate accommodations/access features to individual students, including assigning special test forms; • 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; • 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. **1 – Partially Meets:** 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. **0 – Does Not Meet:** 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. 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.	

A.5.2.1	Generalizability	The assessment program offers appropriate accommodations/ access features that address the access needs of the vast 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.	**2 – Meets:** The assessment program has documentation that is strong and comprehensive regarding the accommodations/access features the program offers, including: • 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; • 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; • 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.); • 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.	

				1 – Partially Meets: 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. **0 – Does Not Meet:** 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 not adhere to its policies and procedures regarding accessibility. 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	
A.5.2.2	Outcomes	The assessment program offers appropriate accommodations/ access features that address the access needs of the vast 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.	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. 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	**2 – Meets:** 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. **1 – Partially Meets:** 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. **0 – Does Not Meet:** The Accessibility Exemplars and accompanying documentation provided by the assessment program indicates neither adequate coverage of major access/ accommodations needs nor adequate quality. 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.	

			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.		
A.5.3	Generaliz-ability	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.).	**2 – Meets:** 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. **1 – Partially Meets:** 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. **0 – Does Not Meet:** Documentation indicates the program "Does Not Meet" at least A.5.1 or A.5.2 regarding English learners. 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.	
A.5.4	Generaliz-ability	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.).	**2 – Meets:** Documentation indicates the assessment program "Meets" both (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. **1 – Partially Meets:** 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. **0 – Does Not Meet:** Documentation indicates the program "Does Not Meet" at least A.5.1 or A.5.2 regarding students with disabilities.	

				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.	

A.6 Ensuring transparency of test design and expectations					
	Type	**Evidence Descriptors**	**Location of Evidence**	**Scoring Guidance**	**Tentative Cut-Offs**
A.6.1	Generaliz-ability	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.	Documentation submitted by assessment program.	**2 – Meets:** All of the following information is available in public documentation that is accurate and organized in a way to be accessible to stakeholders such as policy makers, state assessment program administrators, educators, and parents, and of sufficient quality to promote accurate understanding and uses of the assessments. • 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. • 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 to ensure information will remain current) across every grade level and content area. • Released items are operational items, with annotations and answer rationales provided, including scoring rubrics for constructed-response items with sample responses are provided for each level of the rubric OR the program can demonstrate that they have provided items of operational quality and associated materials that will provide the same or higher levels of information to stakeholders. • Item development specifications are provided. **1 – Partially Meets:** Some of the designated information is not available in public documentation, or information is available but of limited detail or some of the information is inaccurate or inaccessible to stakeholders. Some ways	

				information might be practically inaccessible to public stakeholders include requiring the user to compile information from across multiple documents to yield the information designated above; having information not specifically identified (e.g., having information in a table in a report that is not labeled or searchable for the designated information); not including sufficient information to interpret correctly (e.g., not clearly explaining notation or abbreviations; not clearly including significant exceptions with the information public stakeholders are likely to rely on), etc.0 – Does Not Meet: Large portions of the designated information are not available in public documentation (e.g., two or more bullets are not complete), or large portions are inaccurate and/or inaccessible to stakeholders. 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.	

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	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. Goals include: • In elementary grades, at least three-quarters of the points in each grade align exclusively to the major work of the grade; • In middle school grades, at least two-thirds of the points in each grade align exclusively to the major work of the grade; and • 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. Note: "Major work of the grade" is based on the shifts outlined in the introduction to the CCSS (http://www.corestandards.org/other-resources/key-shifts-in-mathematics/) and described in the K-8 Publisher's Criteria on page 8	Evidence: Test forms, meta-data Specific metadata from assessment program: • Point value of item • Assigned CCSSM alignment (multiple standards shown, if applicable) Coding Sheets: • Do you agree with the assigned alignment? (Y/N) • Revised alignment (if needed) • Does the item align to Major Work? (N/Major) • For High School, does the item align to widely applicable prerequisites? (N/ Prerequisite) Metrics Auto-Calculated: • Number of items • Number and percent of points focused on Major Work. • Number and percent of points focused on not-Major Work. • Number of aligned items. • Percent alignment agreement. • Number and percent of Major Work clusters.	Calculate the percentage of score points that assess the most important content. Assign a score and provide notes under Comments (for each form): For Elementary School: **2 –Meets:** 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. **1 – Partially Meets:** 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. **0 – Does Not Meet:** 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. For Middle School: **2 –Meets:** 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. **1 – Partially Meets:** 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. **0 – Does Not Meet:** 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.	For Elementary School: **2 –Meets:** 75-100% of score points align exclusively to Major Work and at least 90% of the Major Work clusters are assessed **1 – Partially Meets:** 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 **0 – Does Not Meet:** 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. For Middle School: **2 –Meets:** 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. **1 – Partially Meets:** 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. **0 – Does Not Meet:** 0-49% of score points align to the

		(http://www.corestandards.org/wp-content/uploads/Math_Publishers_Criteria_K-8_Spring_2013_FINAL1.pdf), which links to http://www.achievethecore.org/downloads/Math%20Shifts%20and%20Major%20Work%20of%20Grade.pdf showing cluster emphases in footnote 10. "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. (http://www.corestandards.org/assets/Math_Publishers_Criteria_HS_Spring%202013_FINAL.pdf)		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. **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. **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. 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. 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.	Major Work and/or less than 75% of the Major Work clusters for the grade are assessed. 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. **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 **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. 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.

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	Generaliz-ability	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. Goals include: • In elementary grades, at least three-quarters of the points in each grade align exclusively to the Major Work of the grade; • In middle school grades, at least two-thirds of the points in each grade align exclusively to the Major Work of the grade; and • 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.	Evidence: Test blueprints and/or other documents identified by the program.	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: For Elementary School: **2 –Meets:** 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. **1 – Partially Meets:** 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. **0 – Does Not Meet:** 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. For Middle School: **2 –Meets:** 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. **1 – Partially Meets:** 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	For Elementary School: **2 –Meets:** 75-100% of score points align exclusively to Major Work and at least 90% of the Major Work clusters are assessed **1 – Partially Meets:** 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 **0 – Does Not Meet:** 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. For Middle School: **2 –Meets:** 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. **1 – Partially Meets:** 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. **0 – Does Not Meet:** 0-49% of score points align to the

				majority of the Major Work clusters for the grade is assessed. **0 – Does Not Meet:** 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. For High School: **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. **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. **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. 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	Major Work and/or less than 75% of the Major Work clusters for the grade are assessed. For High School: **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. **1 – Partially Meets:** 40-50% of the score points align exclusively to the Major Work and at least 75% of the domains are assessed **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. 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.

				on the domains relevant to the course. 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.	
C.1.3	Generaliz-ability	The assessment design reflects the state’s standards and reflects a coherent progression of mathematics content from grade to grade and course to course.	Evidence: Test blueprints and/or other documents identified by the program.	Assign a score and provide notes under Comments. **2 – Meets:** The test blueprints or other documents indicate that all or nearly all items aligned to the domains listed below reflect adherence to the Progression Documents for the major work of the grade. **1 – Partially Meets:** The test blueprints or other documents indicate that at least three-quarters of the items aligned to the domains listed below reflect adherence to the Progression Documents for the major work of the grade. **0 – Does Not Meet:** The test blueprints or other documents indicate that less than three-quarters of the items aligned to the domains listed below reflect adherence to the Progression Documents for the major work of the grade. Note: Determine that items reflect these Progression Documents: Counting and Cardinality and Operations and Algebraic Thinking (K-5), Expressions and Equations (6-8), and Algebra (HS). Progressions Documents are available at: ime.math.arizona.edu/progressions	**2 –Meets:** 90-100% of the items are aligned to the domains reflecting the Progression Documents for the major work of the grade. **1 – Partially Meets:** 75-89% of the items are aligned to the domains reflecting the Progression Documents for the major work of the grade. **0 – Does Not Meet:** 0-74% of the items are aligned to the domains reflecting the Progression Documents for the major work of the grade.

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	The distribution of score points reflects a balance of mathematical concepts, procedures/fluency, and applications. Goals include at least one-quarter of the points come from each of the following categories: • Conceptual understanding problems in which students to respond to well-designed conceptual problems; • Procedural skill and fluency problems (e.g., purely procedural problems, some requiring use of efficient algorithms, and others inviting opportunistic strategies); and • 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).	Evidence: Test forms, meta-data Specific metadata from assessment program: • Point value of item • Assigned CCSSM alignment (multiple standards shown, if applicable) Coding Sheets: • What does the item assess? • Conceptual understanding, • Procedural skill and fluency, • Application, • Combined Metrics Auto-Calculated: • Number and percent of points for conceptual understanding, procedural skill and fluency, application, and combined (separate categories).	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): **2 –Meets:** At least one quarter and no more than half of the score points are allocated for EACH of the three categories: • Conceptual understanding; • Procedural skill and fluency; and • Application. **1 – Partially Meets:** less than one-quarter of the score points are allocated for one or more of the above three categories. **0 – Does Not Meet:** much less than one-quarter of score points are allocated for one or more of the above three categories. 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.	**2 –Meets:** 25-50% are allocated for each of the three categories **1 – Partially Meets:** 19-24% of score points are allocated for one of the three categories **0 – Does Not Meet:** Less than 18% of the score points are allocated for one or more of the three categories

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.2.2	Generaliz-ability	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.	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: **2 –Meets:** 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: • Conceptual understanding; • Procedural skill and fluency; and • Application. **1 – Partially Meets:** 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. **0 – Does Not Meet:** 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. 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.	**2 –Meets:** 25-50% are allocated for each of the three categories **1 – Partially Meets:** 19-24% of score points are allocated for one of the three categories **0 – Does Not Meet:** Less than 18% of the score points are allocated for one of the three categories

C.2.3	Generaliz-ability	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.	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: **2 –Meets:** Documentation indicates that all or nearly all forms balance conceptual understanding, procedural skill and fluency, and application at all performance levels. **1 – Partially Meets:** Documentation indicates that most, but not all, all forms balance conceptual understanding, procedural skill and fluency, and application at all performance levels. **0 – Does Not Meet:** Documentation indicates that many forms will not balance conceptual understanding, procedural skill and fluency, and application. 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.	**Meets:** 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. **Partially Meets:** 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). **Does Not Meet:** Fewer than 75% of students will be given a form that Meets C.2.2 (score of 2)
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	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).	Evidence: Test forms, meta-data Specific metadata from assessment program: • Point value of item • Assigned CCSSM alignment (multiple standards shown, if applicable)	Calculate the percentage of items that assess mathematical practices and content. Assign a score and provide notes under Comments (for each form): **2 –Meets:** All or nearly all items that assess mathematical practices also align to one or more content standards. **1 – Partially Meets:** The large majority of items that assess	**2 –Meets:** 90-100% of the items that measure a mathematical practice also align to a content standard. **1 – Partially Meets:** 75-89% of the items that measure a mathematical practice also align to a content standard.

		Goals include: • 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); • Through the grades, test items reflect growing sophistication of mathematical practices with appropriate expectations at each grade level.	Coding Sheets: • If the item measures a mathematical practice, does it align to a content standard? (Y/N) Metrics Auto-Calculated: • Number and percent of items measuring practices that also measure content. • Number and percent of items measuring practices that do not measure content.	mathematical practices also align to one or more content standards. **0 - Does Not Meet:** Less than a large majority of items that assess mathematical practices are aligned to one or more content 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.	**0 – Does Not Meet:** 0-74% of the items that measure a mathematical practice also align to a content standard.

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	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: • Every test item that assesses mathematical practices is also aligned to one or more content	Evidence: Test blueprints and/or other documents identified by the program.	Assign a score and provide notes under Comments. **2 –Meets:** Documentation indicates that all or nearly all items that assess mathematical practices also align to one or more content standards. **1 – Partially Meets:** Documentation indicates that the large majority of items that assess mathematical practices also align to one or more content standards. **0 – Does Not Meet:** Documentation indicates that less than a large majority of items that assess mathematical practices are aligned to one or more content 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.	**2 –Meets:** 90-100% of the items that measure a mathematical practice also align to a content standard. **1 – Partially Meets:** 75-89% of the items that measure a mathematical practice also align to a content standard. **0 – Does Not Meet:** 0-74% of the items that measure a mathematical practice also align to a content standard.

		standards (most often within the Major Work of the grade); • Through the grades, test items reflect growing sophistication of mathematical practices with appropriate expectations at each grade level.			

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.3.3	Generaliz-ability	Goals include: • 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). • Through the grades, test items reflect growing sophistication of mathematical practices with appropriate expectations at each grade level.	Evidence: Test blueprints and/or other documents identified by the program.	Assign a score and provide notes under Comments. **2 –Meets:** The test blueprints or other documents indicate that all or nearly all items that assess mathematical practices also align to one or more content standards AND all or nearly all items reflect growing sophistication of mathematical practices across the grades. **1 – Partially Meets:** The test blueprints or other documents indicate that the large majority of items that assess mathematical practices also align to one or more content standards AND the large majority of items reflect growing sophistication of mathematical practices across the grades **0 - Does Not Meet:** The test blueprints or other documents indicate that less than a large majority of items that assess mathematical practices are aligned to one or more content standards AND less than the large majority of items reflect growing sophistication of mathematical practices across the grades Insufficient information box checked if there is insufficient information to score. Comments must be added to explain rationale for insufficient	**2 –Meets:** 90-100% of the items that measure a mathematical practice also align to a content standard and reflect growing sophistication of mathematical practices across the grades. **1 – Partially Meets:** 75-89% of the items that measure a mathematical practice also align to a content standard and reflect growing sophistication of mathematical practices across the grades. **0 – Does Not Meet:** 0-74% of the items that measure a mathematical practice also align to a content standard and reflect growing sophistication of mathematical practices across the grades.

				information determination. For example, one or more pieces of evidence listed in the "Location of Evidence" column were not available.	
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**
C.4.1	Outcome	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 • Mathematical topic coverage in the task (single topic vs. two topics vs. three topics vs. four or more topics); • Nature of reasoning (none, simple, moderate, complex); • Nature of computation (none, simple numeric, complex numeric or simple symbolic, complex symbolic); • Nature of application (none, routine word problem, non-routine or less well-posed word problem, fuller coverage of the modeling cycle); and	Evidence: Test forms Specific metadata from assessment program: • Point value of item • Assigned CCSS alignment (multiple standards shown, if applicable) • If program uses Webb, assigned item DoK • If program does not use Webb, assigned item cognitive demand level Coding Sheets: • By Standard: primary DoK, secondary DoK, tertiary DoK, quaternary DoK. • By item: Indicate DoK Metrics Auto-Calculated: For each test form: • Number and percent of standards at each of the DoK levels • 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	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). **2 –Meets:** 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. **1 – Partially Meets:** 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. **0 – Does Not Meet:** 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	**2 – Meets:** • The DoK Index is at least 80% AND • the percentage of score points associated with DoK3+ items is no more than 10% less than the percentage of standards that are DoK3+. **1 – Partially Meets:** • The DoK Index is at least 60% AND • the percent of DoK1 score points is no more than 20% higher than the percentage of standards that are DoK1. **0 – Does Not Meet:** • The DoK Index is less than 60% OR • the percent of DoK1 score points is more than 20% greater than the percentage of standards that are DoK1.

		• Cognitive actions (knowing or remembering, executing, understanding, investigating, or proving).	percentages of the minima • DoK Index averaged across both test forms.	of Evidence” column were not available.	
C.4.2	Generalizability	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 • Mathematical topic coverage in the task (single topic vs. two topics vs. three topics vs. four or more topics); • Nature of reasoning (none, simple, moderate, complex); • Nature of computation (none, simple numeric, complex numeric or simple symbolic, complex symbolic); • Nature of application (none, routine word problem, non-routine or less well-posed word problem, fuller coverage of the modeling cycle); and • Cognitive actions (knowing or remembering, executing, understanding,	Evidence: Test blueprints and/or other documents identified by the program.	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. **2 –Meets:** 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. **1 – Partially Meets:** 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. However, one or more of these pieces of information is inadequately described or justified. 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. **0 – Does Not Meet:** Documentation does not indicate a definition of cognitive demand, a way of operationalizing cognitive demand at the item level, or a	**2 – Meets:** •If the program uses Webb, the DoK Index is at least 80% AND • the percentage of score points associated with DoK3+ items is no more than 10% less than the percentage of standards that are DoK3+. • 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. **1 – Partially Meets:** • If the program uses Webb, the DoK Index is at least 60% AND • the percent of DoK1 score points is no more than 20% higher than the percentage of standards that are DoK1.

		investigating, or proving).		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. 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.	• 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. • However, one or more of these pieces of information is inadequately described or justified. **0 – Does Not Meet:** • If the program uses Webb, the DoK Index is less than 60% OR • the percent of DoK1 score points is more than 20% greater than the percentage of standards that are DoK1. • 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.
C.5: Ensuring high-quality items and a variety of item types: 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 Guidance**	**Tentative Cut-Offs**
C.5.1	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 selected-response, short and extended constructed-response, technology-enhanced, and multi-step problems.	Evidence: Test forms, meta-data Specific metadata from assessment program: • Item type Coding Sheets: • Are there 2 or more item types? (Y/N) • Does at least one of the item types require students to generate, rather than select, a response? (Y/N) Metrics Auto-Calculated: • Number and percent of traditional multiple-choice items. • Number and percent of multi-select items. • Number and percent of evidence-based selected response items. • Number and percent of technology enhanced items (does not require student to generate a response). • Number and percent of constructed responses. • Number and percent of other item type.	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: **2 –Meets:** At least two item formats are used, including one that requires students to generate, rather than select a response (i.e., CR, gridded response). **1 – Partially Meets:** At least two item formats are used but the item formats only require students to select, rather than generate a response. **0 – Does Not Meet:** Only a traditional multiple choice format is used. 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.	**2 –Meets:** At least two item formats are used, including one that requires students to generate, rather than select a response (i.e., CR, gridded response). **1 – Partially Meets:** At least two item formats are used but the item formats only require students to select, rather than generate a response. **0 – Does Not Meet:** Only a traditional multiple choice format is used.

C.5: Ensuring high-quality items and a variety of item types: 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 Guidance	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	Evidence: Test forms, meta-data Specific metadata from assessment program: • Point value of item • Assigned CCSSM alignment (multiple standards shown, if applicable) • Item Type • Keyed Correct Answer • Rubrics for open-ended items Coding Sheets: • Is there a quality issue with this item? (Y/N) • If so, what is the issue? (Select all that apply) - Item may not yield valid evidence of targeted skill - Item has issues with readability - Item incorrectly keyed - Item has unintended correct answers - Mathematically inaccurate Metrics Auto-Calculated: •Number and percent of high-quality items. • Number and percent of points by issue type, combined, & total. • Number and percent of constructed- and fixed-response types. • Number and percent of agreement with given alignment.	Using the test forms and metadata, determine that there are high-quality items. Assign a score and provide notes under Comments: **2 –Meets:** Nearly all operational items reviewed reflect technical quality, alignment to standards, and editorial accuracy. **1 – Partially Meets:** 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. **0 – Does Not Meet:** Several operational items reviewed have issues with technical quality, alignment to standards, and/or editorial accuracy. 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.	**2 – Meets:** 95-100% for editorial and technical; 90% for alignment to standards **1 – Partially Meets:** 90-94% for editorial and technical; 80% for alignment to standards **0 – Does Not Meet:** 0-89% for editorial and technical; 0-79% for alignment to standards

C.5.3	Generaliz-ability	To support claims of quality, the following are provided in documentation: • Rationales for the use of the specific item types; • Specifications showing the proportion of item types on a form; • 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; • A description of the process used for ensuring the technical quality, alignment to standards, and editorial accuracy of the items.	Evidence: Test blueprints, administration and scoring manuals, QC procedure documents, and/or other documents provided by the program.	Assign a score and provide notes under Comments: **2 –Meets:** Documentation supports claims of the technical quality, alignment to standards, and editorial accuracy. **1 – Partially Meets:** Documentation partially supports claims of the technical quality, alignment to standards, and/or editorial accuracy. **0 – Does Not Meet:** Documentation does not support claims of the technical quality, alignment to standards, and/or editorial accuracy. 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.	**2 –Meets:** Documentation supports claims of the technical quality, alignment to standards, and editorial accuracy. **1 – Partially Meets:** Documentation partially supports claims of the technical quality, alignment to standards, and/or editorial accuracy. **0 – Does Not Meet:** Documentation does not support claims of the technical quality, alignment to standards, and/or editorial accuracy.
C.5.4	Generaliz-ability	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.	Assign a score representing the specification for ensuring high-quality items and a variety of item types; provide notes under Comments: **2 – Meets:** 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). **1 – Partially Meets:** Documentation indicates that at least two formats, but the item formats only require students to select, rather than generate a response. **0 – Does Not Meet:** Documentation indicates that only a traditional multiple choice format is used.	**2 – Meets:** 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). **1 – Partially Meets:** Documentation indicates that at least two formats, but the item formats only require students to select, rather than generate a response.

				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	**0 – Does Not Meet:** Documentation indicates that only a traditional multiple choice format is used.

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
			❑: Missing			
		Comment:				
		A.5.1.2				
			❑: Missing			
		Comment:				
		A.5.1.3				
			❑: Missing			
		Comment:				
A.5.2	**Offering appropriate accommodations/access features**	A.5.2.1				
			❑: 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		
			❑: Missing			
		Comment:				
A.5.3	**English learners**	A.5.3				
			❑: Missing			
		Comment:				
A.5.4	**Students with disabilities**	A.5.4				
			❑: Missing			
		Comment:				

Criterion		Sub-Criterion	Score	Group Criterion-Level Raw Score	Group Criterion Score Rules
A.6.1	**Information available to the public**	**A.6.1**	(0/1/2 Score)		E G L W
			❑: Missing		
		Comment:			

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			
		Comments:								
		C.1.3			(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			
					❑: 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	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.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			
					❏: Missing					
		C.2.3			(0/1/2) Rating					
					❏: 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
			❏: Missing	❏: Missing			❏: Missing	❏: 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 ❏: Documentation missing			
					❏: Missing					
		Comments:								
		C.3.3			(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			
					❏: 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	❑: Missing	❑: 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	❑: Missing	❑: Missing	Add (0/1/2) ratings from each form and each outcome sub-criterion. Range: 0 to 4	E G L W
		Comments:								
		C.4.2			(0/1/2) Rating ❑: 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			
		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	❑: Missing	❑: 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	❑: Missing	❑: 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	❑: Missing	❑: Missing			❑: Missing	❑: Missing		
		Comments:								
		C.5.3			(0/1/2) Rating ❑: 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			
		Comments:								
		C.5.4			(0/1/2) Rating ❑: 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			
		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, 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 B: EVALUATION OF COGNITIVE DEMAND

The CCSSO Criteria ask that the distribution of cognitive demand for each grade level and content area be sufficient to assess the depth and complexity of the standards (Criteria B.4 and C.4). Determining whether these Criteria are met requires four main activities:

1. Coding the content standards to determine what the target distributions of cognitive demand ought to be;
2. Coding the assessment items to determine what the distribution of cognitive demand is for the assessment test form(s);
3. Evaluating the observed cognitive demand of the assessment items in relation to the target cognitive demand of the content standards;
4. Evaluating the intended cognitive demand of the assessment program, as specified in documentation such as test specifications, in relation to the target cognitive demand of the content standards.

While assessment programs may have created their own measures of cognitive demand as part of their test development process, for an external evaluation, especially one of multiple assessment programs, it is useful to have a common way of determining cognitive demand across the programs. Thus, this methodology suggests using the well-regarded and well-known Depth of Knowledge (DoK) measure developed by Norman Webb as the indicator of cognitive complexity. It should be noted that the CCSSO Criteria identify Webb's DoK as an appropriate taxonomy, but suggest that it is preferable to have "classifications specific to the discipline and drawn from the requirements of the standards themselves and item response modes." In the future, such approaches may be developed and incorporated into this methodology.

Coding the Cognitive Demand of the Content Standards

In coding the cognitive demands of the target content standards, the Implementer may rely on previous coding of the standards done by reputable and knowledgeable experts using Webb's DoK levels. Alternatively, the Implementer may have a subset of evaluators conduct this coding. The evaluators will confer about their DoK codings and agree on codings and rationales, which might include multiple DoK for a single standard. The results are averaged across all grade-level standards to determine the proportion at each DOK level. Each standard is equally weighted, because there is no clear indication in the CCSS that any standard is more important than any other. In this analysis, the target content standards will be the state standards that the assessment under review is used to assess.

Coding the assessment items to determine what the distribution of cognitive demand is for the assessment test form(s);

To evaluate the test items, reviewers use the Webb Depth of Knowledge (DoK) framework specific to the content area being evaluated. Each item is rated on the DOK framework; items may be placed into one or two of the four available levels. Based on the item-level ratings, the DOK distribution of the entire test is calculated by averaging across items. To ensure an accurate calculation, test items are weighted by the number of score points associated with each item (e.g., on a two-item test where the first item is a one-point item placed at DOK 1 and the second item is a two-point item placed at DOK 3, the DOK distribution for the test would be 33% DOK 1 and 67% DOK 3).

Evaluating the observed cognitive demand of the assessment items in relation to the target cognitive demand of the content standards;

To reach a score, the DOK distribution for the test form is compared to the recommended DOK distribution for the grade-level standards. The comparison between the DOK distribution of the test and that of the standards is based on two measures. First, the DOK distributions are compared by creating a DOK index, which is based on the proportional agreement between the test form and the standards in DOK distribution (Porter, 2002). Mathematically, this is calculated as the sum of the cell-by-cell minima between the two documents. For example, suppose the standards were coded as being 25% at each of the four DOK level, and the test was coded as being 40% at DOK 1, 40% at DOK 2, and 20% at DOK 3. The DOK index would be .70 (25% from DOK 1, 25% from DOK 2, and 20% from DOK 3).

	Standards	Test	Minimum
DOK 1	.25	.40	.25
DOK 2	.25	.40	.25
DOK 3	.25	.20	.20
DOK 4	.25	0	0
			sum = .70

Second, because a key problem of prior-generation assessments was their low overall DOK, the test is compared with the standards specifically on coverage of higher-level (3+) DOK, with the goal of ensuring that the proportion of the test on DOK 3+ is not markedly lower than that of the standards.

Evaluating the intended cognitive demand of the assessment program, as specified in documentation such as test specifications, in relation to the target cognitive demand of the content standards is a crucial aspect of this evaluation.

The Generalizability review for B.4 and C.4 focuses on the extent to which DOK is an explicit part of the test documentation. Is there is 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?

APPENDIX C:
EVALUATING ACCESSIBILITY ACCORDING TO THE CCSSO CRITERIA

Introduction

The CCSSO *Criteria* include Accessibility, which reflects a concern with fairness, one of the fundamental aspects of validity in testing (AERA/APA/NCME, 2014). CCSSO's Accessibility Criterion encompasses what would be considered accommodations and also access features. In the field, an accommodation is a variation in standardization of an item or administration condition that is intended to support more valid score inferences. What constitutes an "access feature" is not as well defined or agreed upon in the field, but typically includes the following:

- Design of items and test administration procedures intended to support valid score inferences such as to reduce the need for accommodations (sometimes referred to as "universal design").
- Variations in standardization that function as accommodations but that are sometimes administered differently than accommodations, e.g., may not require a formal IEP in order to qualify.
- Variations in standardization that are not intended to be accommodations—that is, they are variations in standardization sponsored by the assessment program but are not related to the construct intended to be assessed.

This document focuses on the first two types of access features that are related to reducing construct-irrelevant variance.

The CCSSO criterion for Accessibility (A.6) focuses on the rationale, development, and validation of accommodations and access features provided by the assessment program.[1] The methodology described in this document does not include validation support involving empirical data from operational administration of the test. In this document, evaluators focus on the adequacy of documentation provided by the assessment program; evaluators evaluate a sample of items and associated documentation on a limited basis. A more complete evaluation of the validity of the accessibility of the program's assessments may be conducted as part of the Test Characteristics evaluation when the assessment program has data from operational administrations to analyze.

The panel will consider evidence submitted ahead of time by the assessment program, which will be informed of the CCSSO evaluation criteria. Requested evidence will consist of documentation and exemplars of accommodations/access features. Documentation submitted by the assessment program may vary from program to program, but will have been selected by the program to support the program's claim it met the criteria. The assessment program will have identified specific places in the submitted documentation that address the features to be evaluated. It is expected that documentation may include such things as 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.

Exemplars are intended to provide evaluators more concrete evidence to ground their understanding of the assessment program's handling of accommodations/access features in conjunction with the program's documentation. The assessment program should select sets of exemplars that show how its accommodations/access features and/or item design are fair for test-takers and support valid score interpretations. The assessment program should select one set of exemplars for ELA/literacy and one set for mathematics. Each set should consist of at least 10 but no more than 25 exemplar items. At least five exemplars should be for high incidence usage . Additionally, programs may submit at least one exemplar for each usage that is essential for a particular disability. Accommodation/ access features for both ELL and SWD students should be included. If the assessment program offers different accommodations/access features for different grades being evaluated, then the assessment program should select a set for each grade level.

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 including rationale to support the features, what the accommodation/access feature is, how it supports more valid score interpretations, instructions for administration, and validity evidence.

The following information should be provided for each item/accommodation/access feature to facilitate review by the evaluators:

- content area and grade
- item number (or other way to uniquely identify the thing being reviewed)
- content standard/construct addressed (if applicable)
- what the accommodation/access feature is; how it differs from the non-accommodated/access version
- instructions for administration/use (if needed)
- conditions under which the accommodation/access feature is available;
- process by which the accommodation/access feature is approved to be used by a student,
- why it is fair in relation to the focal construct and intended score interpretation,
- how it relates to the assessment program's documentation on fairness and item specifications,
- any other salient aspect about the exemplar that the assessment program would like evaluators to be aware of.

The information may come from multiple sources that the assessment program should provide.

The CCSSO evaluation criteria for Accessibility are:

Criterion A.5.1: Following the principles of universal design: 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.

Criterion A.5.2: Offering appropriate accommodations and modifications: 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.

Criterion A.5.3: Assessments produce valid and reliable scores for **English learners**.

Criterion A.5.4: Assessments produce valid and reliable scores for **students with disabilities**.

The evaluation of Test Content will result in a "Preliminary Rating" of Accessibility, because the CCSSO Criteria specify evaluation of the degree to which the "Assessment produces valid and reliable scores," which requires evidence from operational administration (and perhaps special studies) that will be considered in the evaluation of Test Characteristics.

Organizing the Evaluation

The Implementer is responsible for several tasks to organize the evaluation of an assessment program in terms of the CCSSO Accessibility criterion:

- Gather materials from the assessment program
- Recruit and train a panel of evaluators
- Organize the materials the evaluators will use
- Compile ratings and associated information from evaluators

Gather Materials

The Implementer is responsible for gathering the materials needed to conduct the evaluation of the assessment program in relation to the CCSSO Accessibility criterion. The materials include documentation related to accessibility/accommodations, generalizability criteria and Exemplar items.

Recruit and Train Evaluators

The Implementer is responsible for recruiting evaluators with the necessary qualifications, and to train the evaluators appropriately in the specific procedures of the evaluation study.

Evaluating evidence and making judgments about accessibility in relation to the CCSSO *Criteria* requires expertise in the subject area/constructs, how to validly address possible challenges to standardization because of population (e.g., students in grades 3-high school, students with disabilities, English learner's interaction with the program's item types and administration procedures (e.g., human-computer interaction), and large-scale assessment (e.g., item development, forms construction, approval and administration protocols, data gathering). Implementers will need to assemble a qualified panel of persons to evaluate the evidence submitted by the assessment program in relation to the CCSSO Criterion on Accessibility. The panel may overlap with the panel evaluating the other aspects of Test Content, or it may be a panel that focuses only on Accessibility. A typical panel would consist of at least 3-4 persons who together have appropriate expertise. Typical areas of expertise would include the construct being assessed (e.g., reading), accommodation needs of special populations (e.g., English learners, students with disabilities), and the accommodations offered by the assessment program (e.g., technology-based accommodations). Because different issues arise for each content discipline, evaluators should consider disciplines separately (e.g., English language arts and mathematics); some members of the evaluation panel might need to be different to reflect the necessary disciplinary expertise. In addition, an assessment program may make different accommodations available for different grades if the construct changes over grades or if the student needs and abilities change . Evaluators should have appropriate expertise not only in the content area but also at the grade level being evaluated.

Implementers may also consider other desirable qualifications of evaluators, such as their credibility, their ability to do what the evaluation study requires (e.g., participate well in a group, or work independently).

Implementers are responsible for ensuring the evaluators are able to do what they are required to do to produce accurate ratings and comments. Accomplishing this would typically involve training on the specific procedures and materials of the evaluation study, as well as some type of monitoring that the evaluators can apply the training in following the procedures and making accurate judgments.

Rating Procedures

The evaluation of Accessibility involves evaluators working individually and then in groups through these steps.

1. Each evaluator, as an individual, rates the Accessibility Sub-criteria related to Generalizability[1] (A.5.1.1, A.5.1.2, A.5.1.3, and A.5.2.1) separately for ELs and SWDs.
2. Each evaluator, as an individual, rates the Accessibility Sub-criteria related to Outcomes[1] (A.5.2.2) separately for ELs and SWDs.
3. Each evaluator reviews all of the EL data collected from the Generalizability and Outcome results and completes the rating for A.5.3.
4. Each evaluator reviews all of the SWD data collected from the Generalizability and Outcome results and completes the rating for A.5.4.
5. Evaluators as a group assign a tentative score for the Generalizability Sub-criteria as a (A.5.1.1, A.5.1.2, A.5.1.3, and A.5.2.1) separately for ELs and SWDs.
6. Evaluators as a group assign a tentative score for the Outcome Subcriterion (A.5.2.2) separately for ELs and SWDs.
7. Evaluators as a group assign a tentative score for A.5.3. and A.5.4.
8. Finally, evaluators as a group consider their ratings of Generalizability and Outcomes together to assign the final rating for the Accessibility Criterion.

<u>Step 1: Individual Rating of Each Sub-Criterion Related to Generalizability</u>

Each evaluator will individually rate the evidence for the first four Accessibility Sub-criteria (A.5.1.1, A.5.1.2, A.5.1.3, and A.5.2.1) related to Generalizability on a 0-2 point scale (Does Not Meet/Partially Meets/Meets), based on documentation provided by the assessment program. Evaluators will rate each separately for ELs and SWDs. They will determine the overall rating for the overall sub-criterion (e.g., A.5.1.1 or A.5.2.1).

<u>Step 2: Rating of Accessibility Outcome Sub-Criterion based on Exemplars</u>

1. Each evaluator individually rates one Accessibility Sub-Criterion, based on Exemplar items and associated documentation provided by the assessment program. The group of evaluators will evaluate each exemplar, along with its associated documentation using the Exemplar Scoring Form. The organization implementing the evaluation may arrange for evaluators to work individually prior to working as a group, but that is not required. Working individually prior to working as a group may be a way to decrease the time the group needs to meet together, which may help with logistical and cost factors. However, the scores and comments should be based on discussion by the group. The group will record its score and associated comments separately for ELs and SWDs. The score and comments should represent the majority of the panel if not full consensus; a minority position should be documented in the comments.

   Since scores are assigned for ELs and SWDs separately, the overall combined rating should not be higher than the lower of these ratings. However, of the group consensus is to assign the higher score, a solid rationale needs to be provided.

2. You, individually or as a group, will record notes about each Exemplar, starting on page 3 of the Exemplar Scoring Form.

    2.1 For each Exemplar, you will record a short description to help you remember and refer to it, if needed.

    2.2 As you examine each Exemplar and its supporting documentation provided by the vendor, you will record the Need(s) Addressed. This note should be specific enough that you can distinguish it from other Exemplars.

    For example, "Provides student-choice of magnification of screen, including fonts and graphics for students with low vision or those who wish to examine graphic details; program provides documentation about 8% of the population needs this feature as an accommodation; 7% were approved as an accommodation; 15% used this feature in 2015 administration" is better than "Provides access for low vision students."

    2.3 As you examine each Exemplar and its supporting documentation, you will record the Quality, including whether it has any major problems that would make it unacceptable for operational use. This note should be specific enough that you can use it to assemble a portrait of the overall quality of the Exemplars at the end. The evaluation panel will need to rely on their professional judgment to evaluate Quality of the Exemplars because it is impossible to determine beforehand exactly what the Exemplars will address. Quality might include such aspects as: Coherence (the Exemplar is matched to the Need), Correctness (the Exemplar provides appropriate accommodations/access), Adequacy (the Exemplar fulfills the Need), Innovation (the Exemplar addresses a Need in a new or insightful way), Execution (the Exemplar works as intended). Each of these could be made more specific. For example, aspects of Execution might include clarity of instructions, ease of use of features for intended population, appropriate language/graphics, proper rendering for computer-based administration, controls all work as intended, etc.

    If the Exemplar is of adequate Quality for an operational test, then mark "Y" under "Quality." If the Exemplar is not of adequate Quality, then mark "N." You must record notes that provide support for any "N" rating.

    2.4 It is recommended that the assessment program submit 5 exemplars that are most widely used in the program. The intent of this is to ensure the collection of exemplars provides some evidence that needs are addressed in terms of frequency of demand. On the coding sheet, circle the number of the exemplar if it is an exemplar identified by the assessment program as one of the 5 most frequently used. Note that frequency of use is important, but there are many other indicators of addressing needs and quality.

3. You as a group will summarize your notes under "Needs Addressed." The purpose of this is to help you, as a group, reflect on the range and depth of Needs Addressed by the set of exemplars. For example, you might find it useful to group together similar Needs Addressed. You should reflect on what the set of exemplars indicates about the program as a whole. Is it comprehensive, or are there substantive gaps? Is it coherent in terms of enacting a thoughtful approach to fairness? Are you confident that the evidence provides a good representation of the program?

    If you identify any substantive Need Not Addressed, record it in the box, and be sure to include an explanation in your notes.

4. You, as a group, will summarize your notes under "Quality." The purpose of this is to help you as a group reflects on the range and depth of Quality indicated by your examination of the set of exemplars. You should discuss and review carefully any exemplar marked as "No" in terms of demonstrating adequate quality. Record any Exemplar marked "No" in the box provided, and be sure to include an explanation in your notes. Your summary should enable you to characterize the Quality of the set of exemplars and also the quality of the program in addressing fairness.

<u>Step 3: Individual Scores for Criteria A.5.3 and A.5.4</u>
These criteria are sub-scores specific to ELs and SWDs, Data captured from Generalizability and Outcome reviews have been captured for ELs and SWDs separately throughout the review. Reviewing all of the previous scores captured for each sub-criterion, scores for these are determined using the scoring criteria.

1. Evaluators will review all of the previous information specified under A.5.1 and A.5.2 and determine a separate score for ELs (A.5.3) and SWDs (A.5.4).

Group Score and Comments
Directions for this rating are found in the Access/accommodations Exemplar Review Instructions and accompanying Exemplar Review Scoring Form.

Group Ratings

Step 3: Group Tentative Rating of Generalizability Sub-Criteria

The evaluators as a group review and discuss their scores and comments for the Generalizability Sub-criteria (5.1.1, 5.1.2, 5.1.3, 5.2.1). Comments should provide information about the basis for the rating, including areas of strength and areas for improvements. Based on their professional judgment, reviewers assign a tentative rating of E, G, L, or W for the assessment program's documentation as a whole in relation to the CCSSO Accessibility Sub-criteria. The group records reasons for their tentative rating.

Some tentative guidance in making this rating is provided below.

1. Consider A.5.1.1, 5.1.2, 5.1.3, and 5.2.1. (5.3 and 5.4 overlap with these.) The maximum total number of points for these four Sub-criteria is 8 points.
2. Use the following guidelines to start your discussion of what rating to assign:
   7-8 points = E (program was rated "2" on at least three out of four, with no more than one "1" rating)
   5-6 points = G (program was rated a "2" on one or two criteria, and no lower than a "1" on the rest)
   3-4 points = L (program was rated an average of "1" on all four, with no more than one "0" rating
   0-2 points = W

Professional judgment would consider the nature and extent of strengths and weaknesses in addition to the number of points. For example, one criterion might be "Partially Meets," but the evaluators might judge the lack so serious that the rating should be G rather than E. Conversely, they might have rated three criteria as "Partially Meets", but when they look at the specific lack, it might be minor enough (the same thing was missing from three criteria) and the other areas so strong that the overall rating should be a G. The group should record Comments to help others understand their rating.

Step 4: Group Tentative Indication of Outcome Sub-Criterion

The evaluators as a group review and discuss their scores and comments for the Outcome Subcriterion. To generate tentative group scores and Comments, reviewers will summarize two aspects of the set of Exemplars: the degree to which they address the accommodation/access Needs of all students in the intended population, and the Quality of the Exemplars. The evaluation panel will need to rely on their professional judgment to evaluate Quality of the Exemplars because it is impossible to determine beforehand exactly what the Exemplars will address. Quality might include such aspects as: Coherence (the Exemplar is matched to the Need), Correctness (the Exemplar provides appropriate accommodations/access), Adequacy (the Exemplar fulfills the Need), Innovation (the Exemplar addresses a Need in a new or insightful way), Execution (the Exemplar works as intended). Each of these could be made more specific. For example, aspects of Execution might include clarity of instructions, ease of use of features for intended population, appropriate language/graphics, proper rendering for computer-based administration, controls all work as intended, etc.

Based on their summaries of Needs Addressed and Quality, the groups of evaluators use their professional judgment to apply the scoring guidance and assign a tentative score of O, 1 or 2 to the set of Exemplars and associated documentation in relation to the CCSSO Accessibility Sub-Criterion. The group notes reasons for their tentative score under Comments.

Step 5: Group Rating of Accessibility Criterion

The evaluators as a group review and discuss their tentative rating for the Generalizability Sub-criteria and their tentative indication for the Outcome Sub-Criterion together. They assign a rating for the Accessibility Criterion of E, G, L, or W. The Comments should include a rationale, such as whether the Generalizability and Outcome results largely reinforced each other in terms of the rating, or whether there were noticeable differences. Comments may also include other important information that goes beyond adequacy, such as particular strengths or suggestions of areas to improve.

The EGLW Group Rating for Accessibility should reflect the dominant judgment of the group, but evaluators do not need to reach consensus. The group should record the EGLW Rating for the Accessibility Criterion on the Group Final Accessibility Criterion Rating Form, along with appropriate Comments.

APPENDIX D:
ASSESSMENT PROGRAMS WITH MANY FORMS: PROCEDURES TO SELECT TEST FORMS AND COMPUTER-BASED SUMMARIES

Form Selection

Most assessment programs will have multiple forms for each assessment, and in the case of computer adaptive assessments, will have very many forms or test events generated. Thus, evaluators must consider how to select the forms that will be subject to review in a manner that ensures the integrity and credibility of the evaluation process and results.

Given the practical time and logistical constraints of mounting an evaluation, reviewing two forms or two test events of each assessment evaluated, should be sufficient basis for an evaluation when coupled with a review of Generalizability documentation.

There are many reasonable approaches to selecting forms for fixed form assessments (i.e., those that are non-adaptive and include a pre-determined, limited number of forms). These include asking assessment programs to submit any two operational, already administered forms or asking programs to submit 4-6 forms and then selecting two randomly from that set. Regardless of approach, the forms should represent the assessment program's blueprints and other specifications, and not be a "special form," e.g., a form designed for students with low vision.

For computer adaptive assessments, there are literally millions of test events that could take place. Thus, how can reviewers examine Outcomes, that is, what was actually experienced by students taking the assessments? We recommend evaluators examine two forms specified to represent a spread in student performance likely to pick up differences, if any, on CCSSO Criteria. For example, one test event could be drawn from the events that were/could be administered to students at the 40th percentile of student achievement, and the other test event could be drawn from the events that were/could be administered to students at the 60th percentile of student achievement. However, selection criteria can be modified based on the interests of the evaluation Sponsor or Implementer. For example, an Implementer focused on how well very high and very low-performing students are assessed may draw test events from events that were administered to students in the 10th and 90th percentile. Regardless of which forms are selected, both test events must have been generated using the operational item selection algorithm.

Summary Information for Review

Assessment programs that have many forms may have available computer-based summaries of information suitable for informing the CCSSO *Criteria* evaluation. An assessment program may capture information of which forms are administered to students as part of a computer-administered program; in particular, computer-adaptive testing programs typically have this capability. An assessment program may also generate information about many forms as part of computer simulations performed to understand the properties of the test, such as technical characteristics of the test form (e.g., test information function) or to check the functioning of aspects such as the adaptive algorithm, item pool, and delivery platform. The assessment program may have also used a simulation or computer-based analysis to generate information for an alignment study.

Two examples may help illustrate how an assessment program may have computer-based summaries of item/test form information that would be helpful for the evaluator review in relation to the CCSSO *Criteria*. One, an assessment program may run a number of simulations to understand better the interactions between its CAT item selection algorithm and its available item pool. A simulation might involve generating a set of test forms for 50,000 hypothetical students with a given ability distribution. That set of 50,000 test forms could be automatically analyzed and summarized into a computer-based summary about the nature of that set of test forms. A second example for an assessment program with many test forms that are administered as fixed forms (not CAT) might be an assessment design that consists of two main sections: the first section contains all multiple-choice format items and the second section contains several constructed-response format items. The assessment program has developed 5 versions of the multiple-choice and 5 versions of the constructed-response sections. It will mix those to produce 25 unique test forms, each with a different pair of multiple-choice and constructed-response sections. The assessment program may have a computer-based tool that analyzes and documents information about the 25 different test forms.

If CAT programs provide computer-based summaries of information suitable for informing the CCSSO *Criteria* evaluation, these summaries are considered as part of the Generalizability sub-criteria review; such evidence should be weighed heavily because it accurately reflects the complete set of test forms and/or items.

Compile Summary Information

1. Determine what information the assessment program has that is specifically required for the CCSSO *Criteria* Test Content evaluation, and whether this information is available at the item-level to be incorporated into computer-generated summaries. If some of the CCSSO *Criteria* evaluation features are not already generated, would the assessment program be willing to generate that information?
2. Determine in what form the summary information is available. The summary information might be primarily in descriptive form, e.g., “This number/percentage of text passages were informational text.” The information might also be in evaluative form, e.g., “This number/percentage of test forms met the criterion for proportion of informational text passages out of total text passages.” Information in evaluative form is faster for an evaluator to use, as long as the criteria used to generate the summary match the evaluation criteria exactly. The evaluator will likely want to be able to disaggregate or trace the summary information to check the accuracy of the summary. For example, if the highest level summary report includes, “On 99% of the test forms, at least 75% of the text passages were informational,” then the evaluator may want the assessment program to provide additional information that identifies specific test forms so the evaluator could check that text passages coded as informational were indeed informational according to the evaluation criteria.

An example partial summary is shown below.

> 1. Subject _____ Grade _____ Year _____
> 2. Number of test forms included in this report _____
> 3. Percent of forms where distribution of passage type (% informational) met the CCSSO *Criteria for B.1.1 _____%* *or*
>   Percent of forms with number/percent of informational text passages *or*
>   Percent of forms that met test blueprint regarding distribution of informational/literary text passages (as long as test blueprint corresponds with CCSSO evaluator criteria)

3. Determine in what format the computer-generated summary information is available (e.g., ideally concisely compiled into a few tables organized by CCSSO subcriterion).
4. The assessment program should provide documentation sufficient for the evaluator to be able to interpret the data, including information on the representativeness of the items/test forms included in the summary, the procedures used to generate the summary, the layout and characteristics of the summary reports, and pertinent definitions and other documentation to allow the evaluators to understand the reports.

APPENDIX E:
ADDITIONAL SUGGESTIONS REGARDING IMPLEMENTATION OF CRITERIA FOR PROCURING AND EVALUATING HIGH-QUALITY ASSESSMENTS (CCSSO, SPRING 2015)

In 2014, CCSSO developed *Criteria for Procuring and Evaluating High-Quality Assessments* (the *Criteria*) as a resource states could consider as they develop procurements and evaluate options for high-quality state summative assessments aligned to college- and career-readiness standards.   After the *Criteria* were developed, the National Center for the Improvement of Education Assessment (the Center) saw value in creating a detailed and comprehensive methodology that could be used by states, research organizations, and others to review the extent to which existing or planned summative assessments meet the *Criteria*.

To inform the development of a methodology that will be useful to states and other stakeholders, CCSSO has supplemented the original *Criteria* with: (1) a summary reporting template for providing assessment review results to state leaders and other stakeholders in a clear and useful format; and (2) guidance on the evidence that might lead to a rating of "meeting," "partially meeting," or "not meeting" the standard for each sub-criterion. This information is attached.

The remainder of this document contains a suggested "Summary Reporting Template" and two scoring templates, one regarding the extent to which ELA/literacy assessments meet the CCSSO *Criteria* and one regarding the same for mathematics.

Summary Reporting Template for Test Content review

Results of Applying the CCSSO Criteria for High-Quality Assessments in Test Content	**Degree of Match with CCSSO Criteria**			
	Weak	**Limited**	**Good**	**Excellent**
A. Meet Overall Assessment Goals and Technical Quality				
•A.5: Providing accessibility to all students, including English learners and students with disabilities (subset of the criterion)	🔴	🟡	🟢	🟢
• A.6: Ensuring transparency of test design and expectations	🔴	🟡	🟢	🟢
B. English Language Arts/Literacy				
I. Assesses the content most needed for College and Career Readiness *[[summary of rationale and other comments]]*	🔴	🟡	🟢	🟢
• B.3: Requiring students to read closely and use evidence from texts	🔴	🟡	🟢	🟢
• B.5: Assessing writing	🔴	🟡	🟢	🟢
• B.6: Emphasizing vocabulary and language skills	🔴	🟡	🟢	🟢
• B.7: Assessing research and inquiry	🔴	🟡	🟢	🟢
• B.8: Assessing speaking and listening (optional) N/A	🔴	🟡	🟢	🟢

Results of Applying the CCSSO Criteria for High-Quality Assessments in Test Content (continued)

	Degree of Match with CCSSO Criteria			
	Weak	Limited	Good	Excellent
II. Assesses the depth that reflect the demands of College and Career Readiness *[[summary of rationale and other comments]]*	🔴	🟡	🟢	🟢
• B.1: Assessing student reading and writing achievement in both ELA and literacy	🔴	🟡	🟢	🟢
• B.2: Focusing on complexity of texts	🔴	🟡	🟢	🟢
• B.4: Requiring a range of cognitive demand	🔴	🟡	🟢	🟢
• B.9: Ensuring high-quality items and a variety of item types	🔴	🟡	🟢	🟢
C. Mathematics				
I. Assesses the content most needed for College and Career Readiness *[[summary of rationale and other comments]]*	🔴	🟡	🟢	🟢
• C.1: Focusing strongly on the content most needed for success in later mathematics	🔴	🟡	🟢	🟢
• C.2: Assessing a balance of concepts, procedures, and applications	🔴	🟡	🟢	🟢
II. Assesses the depth that reflect the demands of College and Career Readiness *[[summary of rationale and other comments]]*	🔴	🟡	🟢	🟢
• 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	🔴	🟡	🟢	🟢

Explanation of summary report template and weighting of criteria

The CCSSO Criteria being evaluated in this review (A.5, A.6, B.1 – B.9, and C.1 – C.5) are rolled up into four reporting categories to help make the results of the evaluation more understandable by the end user. Those categories are: I. Assesses the **content** most needed for College and Career Readiness; II. Assesses the **depth** that reflect the demands of College and Career Readiness; III. **Accessible** to all students; and IV. **Transparency** of test design and expectations. These reporting categories are based on the CCSSO Criteria. The Criteria that are underlined (B.1, B.2, B.3, and B.5 in ELA/Literacy and C.1 and C.3 in mathematics) will be weighted more heavily in determining the overall rating for the roll-up category.

Why weight some criteria more heavily than others?

The criteria selected to be weighted most heavily on the assessments capture what matters most in in preparing students for college and careers.

For literacy, this includes the careful examination of texts, meaning work in reading and writing that centers on texts. Research shows that students must be able to read texts of adequate range (B.1) and complexity (B.2) and emphasizes students reading those texts closely to draw evidence and knowledge from the text (B.3 and B.5). The criteria selected to be weighted most heavily revolve around the complexity and range of the texts that students are asked to read and the kinds of questions students should address as they write about them. If assessments closely align to these four selected criteria, they will embody the skills needed for students on the path to college and career readiness.

For mathematics, this includes focusing on the content that matters most. Focusing on the most important content (C.1) is a research-based element of high-quality assessments. Connecting practices to content (C.3) ensures that when items include aspects of modeling and making mathematical arguments they are still measuring important content. A focused assessment system helps ensure students have the most critical knowledge and skills to prepare them for college and careers.

It is important to note that every criterion is critical and will have an impact on an assessment program's evaluation. The weightings are meant to indicate which criteria drive a section's rating, though each criterion will be taken into account, and each will receive its own rating, ensuring that specific strengths and development areas are clear.