

**A GUIDE TO EVALUATING
COLLEGE- AND CAREER-READY ASSESSMENTS:**
Focus on Test Characteristics

Evaluation Methodology

MARCH 2016

ERIKA HALL, PH.D.
Center for Assessment

SUSAN LYONS, PH.D.
Center for Assessment



ACKNOWLEDGMENTS

Throughout the development of the Test Characteristics Methodology, the authors received support, feedback and recommendations from multiple friends and colleagues. We are deeply indebted to each of them for helping us create the comprehensive materials and procedures represented by this document.

First, we would like to thank our friends at the Center for Assessment who always made themselves available to comment on drafts of the document or play the role of guinea pig in trying out aspects of the methodology. We are grateful to your contribution and for responding to our multiple requests for support. A special thank you goes out to Thanos Patelis for helping to manage the project and keep us on track.

Second, we would like to thank the myriad of technical experts who reviewed the methodology and provided us with detailed recommendations for improvement, including: Ric Leucht, Suzanne Lane, Laurie Wise, Cathy Welch, Mike Russell, Linda Cook and Joe Ryan, who had the daunting task of evaluating the first iteration of the Criteria Evaluation Framework as well as multiple versions of the complete methodology. Invaluable feedback was also provided by assessment experts supporting PARCC, Smarter Balanced and ACT who greatly informed our thinking on issues related to feasibility and consistency of implementation.

Finally, we would like to thank the High Quality Assessment Project for supporting the development of this methodology; Judy Wurtzel and Joanne Weiss for their patience and thoughtful feedback throughout the development process; and representatives from the Council of Chief State School Officers, including Scott Norton and Kris Ellington.

Thanks to you all!

ACKNOWLEDGMENTS

TABLE OF CONTENTS

INTRODUCTION	5
PART 1: DEVELOPMENTAL CONSIDERATIONS FOR THE TEST CHARACTERISTICS METHODOLOGY	6
• Elaboration of the CCSSO Assessment Quality Criteria.....	6
• Development of a Criterion Evaluation Framework (CEF) to Support the Evaluation of Assessment Quality.....	9
• Brief Overview of the Development and Review Process	10
• Assumptions Underlying the Design of the Evaluation Methodology	11
• A Framework for Operationalizing the CCSSO Criteria.....	11
• Review of Assessment System Outputs	14
• Relationship between the Framework and Original CCSSO Documentation	16
PART 2: GUIDE FOR IMPLEMENTATION	18
• Summary of the Test Characteristics Evaluation Methodology...	18
• Phase 1: Preparation.....	19
- <i>Submission and Organization of Evidence Necessary to Support Review</i>	19
- <i>Identification of the Evaluation Team</i>	20
- <i>Secure Assembly of Evaluation Materials</i>	22
• Phase 2: Evaluator Training and Independent Evaluator Review.....	22
- <i>Evaluation Team Training</i>	22
- <i>Independent Evaluator Review</i>	30
• Phase 3: Group Discussion and Summary of Evaluation Findings.....	30
• Phase 4: Report Generation and Approval	32
• Considerations Related to the Implementation of the Methodology for Assessments Developed for Use in Multiple States	32
• Considerations Related to the Implementation of the Methodology for Multiple Assessments Simultaneously	34
CONTACT	35

TABLE OF CONTENTS

TABLE OF CONTENTS *continued*

APPENDICES	36
• Appendix A: Crosswalk between CCSSO Evidence Statements and Criteria Evaluation Framework.....	36
• Appendix B: Complete Structure of Claims for Evaluation of Test Characteristics.....	43
• Appendix C: Examples of Evidence that May be Provided for Evaluation	47
• Appendix D: Guidelines and Templates to Support the Collection and Organization of Evidence.....	50
• Appendix E: Independent Evaluator Rating Sheet Sample	53
• Appendix F: Independent Evaluator Rating Sheet Sample Specific to Criterion A.5.....	58
• Appendix G: Sample Test Characteristics Summary Report Template.....	61

TABLE OF CONTENTS

INTRODUCTION

The Center for Assessment has been funded by the High Quality Assessment Project¹ to develop methodologies and procedures to help guide persons and organizations who want to apply the Council of Chief State School Officers (CCSSO) Criteria for High Quality Assessments² in evaluating summative assessments designed to measure college and career readiness standards.³ This includes assessments that may not specifically define “college-ready” cut-scores or benchmarks, but were developed specifically to address knowledge and skills defined as necessary for entrance into college, careers and technical education.⁴

Center for Assessment researchers have grouped the CCSSO criteria into two components—those dealing with test content and those dealing with test characteristics and program implementation. The criteria associated with test content focus primarily on the quality of items, the accessibility of item and test content, and the alignment of test content to the priority content of college and career ready standards. The criteria associated with test characteristics focus on the psychometric and statistical properties of assessment instruments and the quality of test administration, reports and supplemental information provided to aid in the interpretation and use of test results.

This Guide to Evaluating College- and Career-Ready Assessments –Focus on Test Characteristics (a.k.a. The Guide) outlines a methodology for evaluating assessments against the CCSSO Criteria related specifically to test characteristics.⁵ It is divided into three parts: Part 1 provides an overview of the CCSSO criteria, outlines key factors influencing the design of the test characteristics methodology, and introduces the Criteria Evaluation Framework (CEF); Part 2 provides detailed guidelines and considerations for implementing the evaluation methodology, and Part 3 consists of Appendices A-F, which provide exemplars, tools and templates designed to further support implementation. The Guide contains a comprehensive summary of the theory, design and procedures underlying the test characteristics methodology and details supporting implementation. Because a large amount of the information provided may not be relevant or consumable for all audiences, we created two companion documents designed to address the needs of different users. The three documents associated with the test characteristics methodology, and their intended users, are as follows:

1. **Guide to Evaluating College- and Career-Ready Assessments –Focus on Test Characteristics** (the current document)–intended to be used primarily by those implementing an evaluation due to the level of detail and specificity provided. However, it serves as invaluable reference to anyone interested in the understanding the process and materials in greater detail after reviewing the Executive Summary.
2. **Executive Summary** – a short document intended to provide a quick, yet comprehensive introduction to the test characteristics evaluation methodology and tools. It includes a summary of Parts 1 and 2 of the evaluation methodology, as well as appendices B and F. The intended audience for the executive summary includes policy makers, potential funders and requesters, and anyone interested in getting a quick snapshot of what the test characteristics methodology entails.
3. **Criteria Evaluation Framework** –intended to be used by: the evaluation team (to support the review and evaluation of evidence provided); those charged with supplying evidence to support evaluation (to better understand the type and scope of documentation expected); and anyone interested in better understanding how the CCSSO criteria were operationalized (e.g., assessment developers, measurement professionals).

Note: Across all three documents, language that is drawn directly from CCSSO’s Criteria for Procuring and Evaluating High Quality Assessments document is shaded in a light green box in order to help differentiate it from the text and materials developed by the Center for Assessment.

¹ The High-Quality Assessment Project (HQAP) 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, Helmsley Charitable Trust, the Charles and Lynn Schusterman Foundation and the William and Flora Hewlett Foundation.

² See the *Criteria for Procuring and Evaluating High Quality Assessments* at the link: <http://www.ccsso.org/Documents/2014/CCSSO%20Criteria%20for%20High%20Quality%20Assessments%2003242014.pdf>

³ It is important to note that the test characteristics methodology as well as the test content methodology were developed independent from the U.S. Department of Education’s Peer Review Guidance released in September of 2015 <https://www2.ed.gov/policy/elsec/guid/assessguid15.pdf>. However, evaluations generated using those methodologies could provide important evidence for submission to peer review.

⁴ Or in earlier grades, to address the pre-requisite skills necessary to meet these expectations by the end of high school.

⁵ Separate documentation for evaluating the criteria dealing with test content, “Guide to Evaluating Assessment Using the CCSSO Criteria for High Quality Assessments: Focus on Test Content” is available at the Center for Assessment’s website, www.nciea.org.

SUMMARY OF PART 1: DEVELOPMENT OF THE TEST CHARACTERISTICS METHODOLOGY

Elaboration of the CCSSO Assessment Quality Criteria

The CCSSO Criteria for High Quality Assessments, which serve as the foundation for this methodology, were developed by the Council of Chief State School Officers to support states as they “develop procurements and evaluate options for high-quality state summative assessments aligned to their college- and career readiness standards.” The CCSSO criteria are grouped into five broad categories:

- | | |
|---|---|
| A. Meet Overall Assessment Goals and Ensure Technical Quality | D. Yield Valuable Reports on Student Progress and Performance |
| B. Align to Standards – English Language Arts/Literacy | E. Adhere to Best Practices in Test Administration |
| C. Align to Standards – Mathematics | F. State Specific Criteria |

Each category contains one or more associated criteria. Each category is identified by a letter, and a criterion is identified by a letter and number. For example, in category A there are seven criteria numbered A.1 through A.7. The Center for Assessment organized the CCSSO Criteria into two groups, one dealing with test content and the other dealing with test characteristics.

The test content evaluation procedures include specific elements of criterion A.5 related to accessibility, criterion A.6 related to transparency of test design and expectations, and all of the criteria associated with categories B and C, related to alignment to standards for ELA and mathematics respectively. This guide, addressing the evaluation of test characteristics, includes criteria A.1-A.4, A.7, technical aspects of criterion A.5⁶, and all of the criteria associated with categories D and E covering reporting and test administration. Criterion category F, state specific criteria, does not include specific criteria and is, therefore, not covered in either document. The criteria addressed in each document are listed in Table 1. While the criteria could have been divided in a number of defensible manners, the researchers at the Center for Assessment believe this division is appropriate given key differences in the skills and expertise necessary to evaluate these two groups of criteria (i.e., content vs. technical), and expectations about when the data/information necessary to support evaluation will be available.⁷

TABLE 1. ORGANIZATION OF CCSSO CRITERIA INTO TEST CONTENT AND TEST CHARACTERISTICS

TEST CONTENT	TEST CHARACTERISTICS
<p>A. Meet Overall Assessment goals and Ensure Technical Quality</p> <ul style="list-style-type: none"> - A.5 Providing accessibility to all students, including English learners and students with disabilities (partial) - A.6. Ensuring transparency of test design expectations. <p>B. Align to Standards - English Language Arts/Literacy</p> <ul style="list-style-type: none"> - 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 and 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 <p>C. Align to Standards - Mathematics</p> <ul style="list-style-type: none"> - 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 	<p>A. Meet Overall Assessment goals and Ensure Technical Quality</p> <ul style="list-style-type: none"> - A.1 Indicating progress toward college and career readiness - A.2 Ensuring that assessments are valid and required for 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 (partial) - A.7 Meeting all requirements for data privacy and ownership <p>D. Yield Valuable Reports on Student Progress and Performance</p> <ul style="list-style-type: none"> - D.1 Focusing on student achievement and progress to readiness - D.2 Providing timely data that inform instruction <p>E. Adhere to Best Practices in Test Administration</p> <ul style="list-style-type: none"> - E.1 Maintaining necessary standardization and ensuring test security

⁶ The specific components of A.5 that are addressed in the test characteristics evaluation are outlined in Table 2.

⁷ A large portion of the information necessary to support a comprehensive evaluation of the test characteristics criteria will not be available until after an assessment has been administered operationally.

A detailed description of each of the CCSSO criteria is provided in the *Criteria for Procuring and Evaluating High Quality Assessments* (referred to from this point on as the CCSSO Criteria document) at the link:

<http://www.ccsso.org/Documents/2014/CCSSO%20Criteria%20for%20High%20Quality%20Assessments%2003242014.pdf>.

The full text associated with each criterion addressed by the test characteristics methodology, taken directly from the CCSSO Criterion document, is provided in Table 2.

TABLE 2. CCSSO CRITERIA ASSOCIATED WITH TEST CHARACTERISTICS EVALUATION*

- A.1 Indicating progress toward college and career readiness:** Scores and performance levels on assessments are mapped to determinations of college and career readiness at the high school level and for other grades to being on track to college and career readiness by the time of high school graduation.
- A.2 Ensuring that assessments are valid for required and intended purposes:** Assessments produce data, including student achievement data and student growth data required under Title I of the Elementary and Secondary Education Act (ESEA) and ESEA Flexibility, that can be used to validly inform the following:
- School effectiveness and improvement;
 - Individual principal and teacher effectiveness for purposes of evaluation and identification of professional development and support needs;
 - Individual student gains and performance; and
 - Other purposes defined by the state.
- A.3 Ensuring that assessments are reliable:** Assessments minimize error that may distort interpretations of results, estimate the magnitude of error, and inform users of its magnitude.
- A.4 Ensuring that assessments are designed and implemented to yield valid and consistent test score interpretations within and across years:**
- Assessment forms yield consistent score meanings within and across years, as well as for various student groups, and delivery mechanisms (e.g., paper, computer, including multiple computer platforms).
 - The score scales facilitate accurate and meaningful inferences about test performance.
- A.5 Providing accessibility to *all* students, including English learners and students with disabilities:**
- Assessments produce valid and reliable scores for English learners
 - Assessments produce valid and reliable scores for students with disabilities.
- A.7 Meeting all requirements for data privacy and ownership:** All assessments must meet federal and state requirements for student privacy, and all data is owned exclusively by the state.
- D.1 Focusing on student achievement and progress to readiness:** Score reports illustrate a student’s progress on the continuum toward college and career readiness, grade by grade, and course by course. Reports stress the most important content, skills, and processes, and how the assessment focuses on them, to show whether or not students are on track to readiness.
- D.2 Providing timely data that inform instruction:** Reports are instructionally valuable, easy to understand by all audiences and delivered in time to provide useful, actionable data to students, parents and teachers.
- E.1 Maintaining necessary standardization and ensuring test security:** in order to ensure the validity, fairness and integrity of state test results, the assessment systems maintain the security of the items and tests as well as the answer documents and related ancillary materials that result from test administration.

In addition to the criterion descriptions listed above, the CCSSO Criteria document lists examples of evidence that should be evaluated to make determinations about quality. The Center for Assessment's Criteria Evaluation Framework includes and expands upon all of the examples provided in the CCSSO Criteria document and provides guidance on how to evaluate these and other sources of evidence.

Importantly, the CCSSO Criteria document also highlights the need for reviewers to account for an assessment's current stage of development when evaluating the type and amount of evidence submitted for consideration. Specifically, the document differentiates between assessments at three different phases of development, as summarized in Table 3.

TABLE 3. EVIDENCE EXPECTATIONS FOR ASSESSMENTS AT DIFFERENT STAGES OF DEVELOPMENT

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

Considerations related to an assessment’s phase of development and how it should be considered in the review and evaluation of evidence, are addressed in Part 2 of this document. Influential factors are highlighted both in the training of evaluators and the guidelines for identifying appropriate types and sources of evidence to support evaluation. For example, when evaluating evidence related to the development of performance level descriptors (PLDs)⁸ for a *pre-existing assessment* a technical manual describing the process implemented, materials reviewed, and the participants in the review would be critical. However, for a test that is to be *newly created*, evidence related to this criterion may be drawn from the procedures proposed by the vendor in a request for proposal (RFP) or a final PLD report developed by the vendor for an assessment program similar to that under review.

As part of the evaluation process, those reviewing evidence will be asked to consider the extent to which the range and quality of evidence provided meets that which would be expected given the assessments current phase of development. For this reason, a newly proposed assessment that provides detailed evidence to support proposed procedures and methodologies may be rated higher than an existing assessment that provides evidence inconsistent with that which would be expected given its years of implementation. If one is interested in comparing evaluation results for multiple assessments this is one of the many factors that will need to be considered. It is also one of the main reasons that consideration of evaluation ratings in the absence of contextual information is discouraged.

⁸ In many states these are referred to as Achievement Level Descriptors (ALDs). For the sake of clarity and consistency, throughout this document they are henceforth only referred to as Performance Level Descriptors (PLDs).

Development of a Criterion Evaluation Framework (CEF) to Support the Evaluation of Assessment Quality

The CCSSO criteria provide a strong foundation upon which to build an evaluation of assessment quality; however, additional detail and structure was necessary to develop a comprehensive evaluation methodology.

For example, consider the examples of evidence associated with criterion D.2, as presented in the CCSSO Criteria document (see Table 4):

TABLE 4. CRITERION D.2

CRITERIA	EVIDENCE
D.2 Providing timely data that inform instruction: Reports are instructionally valuable, easy to understand by all audiences and delivered in time to provide useful, actionable data to students, parents and teachers.	<ul style="list-style-type: none">• A timeline and other evidence are provided to show when assessment results will be available for each report.• A description is provided of the process and technology that will be used to issue reports in as timely a manner as possible.• Evidence, including results of user testing, is provided to demonstrate the utility of the reports for each intended audience.

In relation to this criterion, one might ask, specifically, “What conditions must hold in order for a report to provide timely, instructionally valuable data to users?” And then as a follow up, “What evidence can be collected to show such conditions hold?” and ultimately, “What is needed for the evidence to show sufficiency in meeting the conditions?” It is thinking through questions such as these that allow one to identify the full range of evidence that is necessary and appropriate to support evaluation. To continue with the example above, in examining the three bullets provided by the CCSSO Criteria document, it could also be argued that directions for accessing and viewing score reports must be clear and readily available to users in order for score reports to provide timely and useful information. While it is stated in the Evidence column for D.2 that report timelines and evidence demonstrating the utility of the reports for the intended audience should be provided, what specifically that evidence should look like or demonstrate is not discussed. Is it enough that a timeline for reporting exists, or does it need to meet some minimum requirements to support the stated criteria; specifically, that reports will be delivered to users in a timely manner? Similarly, what does evidence demonstrating the utility of reports for different audiences look like? What information or results should be provided to support claims that reports are instructionally valuable or actionable? To support evaluation, these expectations must be articulated.

The Criteria Evaluation Framework (CEF) expands upon the CCSSO Criteria by: 1) specifying the claims underlying each criterion, 2) describing what sufficient evidence should look like, 3) providing comments and examples that inform the evaluation process, 4) highlighting key connections among claims and criteria, and 5) supporting the credibility of the evaluation by aligning each criterion to the joint *Standards for Educational and Psychological Testing (2014)*, as recommended in the CCSSO Criteria.

The claims presented in the CEF are logical extensions of the criteria statements presented in the original CCSSO document. As a set, the claims provide an efficient way to organize the broad range of evidence necessary to evaluate each criterion resulting in a better-defined and more manageable evaluation process. For example, the overall criterion statement for D.2 is multi-part and can be broken down into at least three key claims:

- D.2.1. Directions for accessing and viewing score reports (when necessary) are broadly distributed and clear to end-users.
- D.2.2. Reporting timelines, procedures and technology provide for the dissemination of test results in a timely fashion.
- D.2.3. The content and structure of score reports provide useful and actionable information for making instructional decisions.

The test characteristics methodology (articulated in Part 2) asks evaluators to make a judgment about the degree to which the evidence supplied supports each claim individually before making a holistic criterion-level evaluation. To facilitate this process evidence is aligned to specific claims rather than criteria. Appendix B shows the breakout of each of the test characteristics criteria into its set of distinct claims.

While the CCSSO Criteria alone are informative, to implement a coherent evaluation and support those responsible for identifying and assembling evidence, the level of detail and transparency reflected in the CEF is necessary. To further support evaluation the Criteria Evaluation Framework provides examples, comments and descriptors intended to aid evaluators in

making decisions regarding the sufficiency of evidence provided for each claim. These additional features are discussed in detail in the section of this report entitled “A Framework for Operationalizing the CCSSO Criteria.”

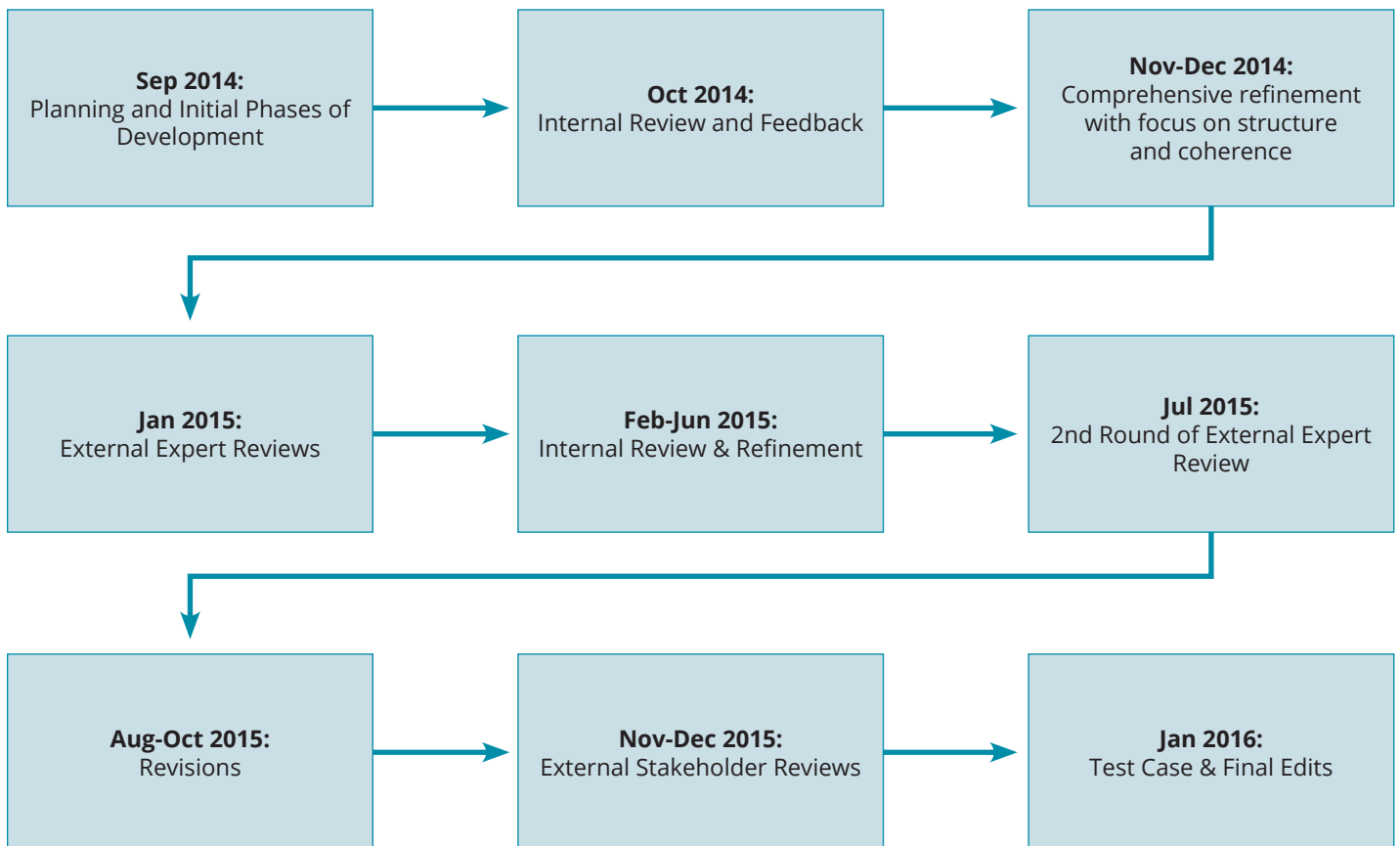
Center for Assessment researchers felt it was important to recognize key relationships across criteria by highlighting how evidence evaluated within the context of one criterion may influence decisions about another. Therefore, while trying to avoid redundancy as much as possible, the Criteria Evaluation Framework identifies critical connections among the criteria through the use of secondary claims. The purpose of secondary claims is to aid reviewers in appropriately evaluating the evidence presented for one criterion in light of all other relevant evidence. For example, Criterion A.2 focuses on the evaluation of evidence supporting the validity of intended inferences. Although not explicitly linked within the context of the CCSSO Criteria document, most would agree that decisions about validity cannot be made without considering the other criteria. Therefore, for A.2, *all* of the claims associated with the other criteria are identified as secondary claims.

Finally, while the CCSSO Criteria were “developed in reference to the Standards for Educational and Psychological Testing (i.e., the *Standards*)” an explicit connection to the Standards was not documented. To enhance the credibility of the methodology and provide additional information to evaluators who wish to use the *Standards* as a companion resource, the Criteria Evaluation Framework identifies the joint standards associated with each criterion (by number only, so as to avoid any copyright issues) making this connection clear.

Brief Overview of the Development and Review Process

Throughout the development of the methodology and the Criteria Evaluation Framework, the materials associated with this effort underwent a series of internal and external reviews, including the implementation of a test case. Figure 1 below broadly summarizes the steps in the development process.

FIGURE 1. DEVELOPMENT TIMELINE



As shown in Figure 1, the development process was iterative in nature, and allowed for the influence of ideas from a number of differing and valuable perspectives. The resulting methodology and associated Criteria Evaluation Framework is much stronger due to the thoughtful and valuable contributions of many.

Assumptions Underlying the Design of the Evaluation Methodology

To establish a coherent evaluation methodology, the primary goal(s) of evaluation, intended user(s), and manner in which results are intended to be used must be transparent and clearly articulated. Some of the key assumptions underlying the design of the test characteristics evaluation methodology and proposed plans for reporting evaluation results are called out in Table 5. These assumptions are addressed in greater detail, where appropriate, throughout the remainder of this document.

TABLE 5. KEY DESIGN ASSUMPTIONS AND CONSTRAINTS.

	KEY ASSUMPTIONS AND CONSTRAINTS
Goals of Evaluation	<p>Provide clear, useful information about the how to evaluate the evidence submitted in support of each of the CCSSO assessment quality criteria so intended users can make informed decisions regarding assessment selection, retention or modification and/or identify areas where additional evidence/research is needed to support claims of overall assessment quality.</p> <p>Establish a comprehensive, yet efficient evaluation process (i.e., with respect to time, resources, capacity, cost, etc.) that provides users with timely, useful and complete information.</p>
Users	<p><i>Intended User of Assessment Evaluation Results</i> – those charged with making decisions about the quality and appropriateness of summative assessments (and/or assessment proposals) used to evaluate student performance relative to college and career readiness standards.</p> <p><i>Secondary User of Assessment Evaluation Results</i> – Test developers/vendors, policymakers outside the sponsoring agency</p>
Evaluators	In order to conduct a valid implementation of the evaluation framework, and provide the type of feedback requested those involved in reviewing and making judgments about evidence will need to have significant, demonstrated psychometric and technical expertise.
Constraints	<p>The CCSSO Criteria, as represented and articulated in Tables 1 and 2, is the basis for the evaluation design.</p> <p>For each criterion, the methodology must result in a holistic rating regarding the overall quality of the assessment based on the evidence provided.</p> <p>The methodology is focused on the evaluation of summative assessments developed to measure college and career readiness standards.</p> <p>The evaluation relies solely on the evidence submitted by the Provider.</p>
Primary References	Standards for Educational and Psychological Testing (AERA, APA, and NCME, 2014); CCSSO Criteria for Procuring and Evaluating High Quality Assessments (CCSSO, 2014)

While the results of evaluation are intended to support users in making informed decisions about the relative quality of different assessments for the specified purpose(s) (e.g., procurement decisions), comparisons should not be based on a single, isolated rating. Assessment programs are designed to meet uniquely specified goals and uses in consideration of different views and philosophies related to the assessment of student proficiency and progress. The test characteristics methodology is designed to support evaluation of the sufficiency of a submitted body of evidence in *light of these contextual factors*. The methodology outlines procedures and materials that provide for qualitative feedback and contextualized ratings about the sufficiency of evidence provided for each criterion, and intentionally *does not* provide for an overall assessment-level score or rating. If an overall rating is desired, for practical and/or policy reasons, the process used to combine results across criteria should be clearly defined by the sponsoring agency in consideration of the overarching goals and priorities for the assessment in question and a clear understanding of the contextual factors at play. In addition, the rationale for the procedures used, and guidelines to support interpretation in light of those procedures should be clearly articulated and included in the final report.

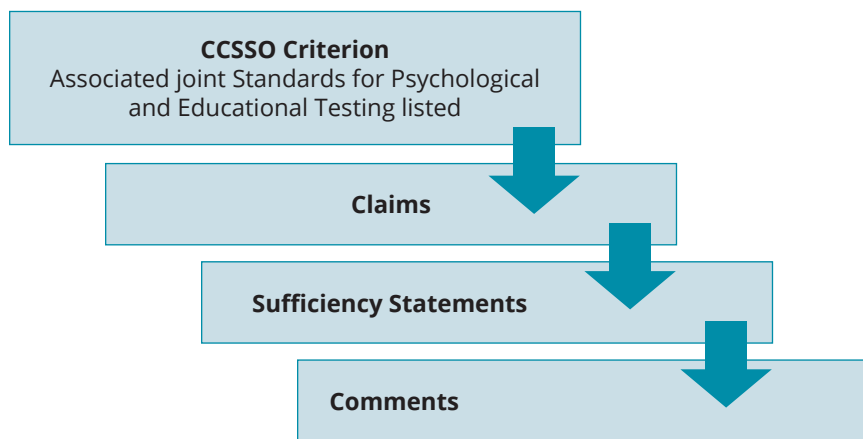
A Framework for Operationalizing the CCSSO Criteria

To guide the evaluation of assessment quality relative to CCSSO's Assessment Quality Criteria, one must not only identify the type of evidence necessary to support evaluation, but also decide what specifically about that evidence should be considered in evaluating each criterion. What constitutes appropriate or high quality evidence will depend on a variety of factors that are likely to vary across assessment programs due to contextual factors. These factors include the manner in which results are to be used,

the stakes associated with those uses (for individuals and systems), and the sponsoring agency⁹ or vendor’s theory of action as to how the assessment will bring about desired change. For many programs other contextual factors, including the “age” of the assessment and degree to which the associated standards have been addressed in the classroom will also be influential. It is for this reason that there can never be a “universal” evaluation system that can explicitly dictate how evidence should be weighed and evaluated for all tests. Instead, the Criteria Evaluation Framework (CEF) provides the structure and content necessary to implement a thoughtful, evidence- and discussion-based evaluation of assessments designed to measure student achievement and progress relative to college- and career-ready standards. While we attempt to operationalize the criteria in a manner that supports a fair, transparent and reliable evaluation, due to the contextual factors previously discussed, expert discussion and judgment will play a significant role in the end result of the evaluation. As the Criteria Evaluation Framework is applied in practice, examples and guidance can continue to be developed to aid stakeholders in considering the array of factors that may influence judgments about the sufficiency of evidence provided for a given assessment.

The CEF was organized around the following hierarchical structure.

FIGURE 2. STRUCTURE OF THE CRITERIA EVALUATION FRAMEWORK (CEF)



In establishing this structure the first step was to articulate a set of **Claims** supporting each CCSSO criterion.

Claims are statements we want to make about procedures, materials, reports, and/or data given the evidence provided for review.

For each criterion there may be any number of associated claims, and as a set, claims suggest not only the type/range of evidence expected but what features of that evidence are important relative to a given criterion.

Claims that were developed in direct consideration of a particular criterion are referred to as *primary claims*. Claims that are primary for one criterion, but relevant to the evaluation of another are listed as *secondary claims* for that secondary criterion. For example, there are 11 primary claims associated with A.1 as shown in Table 6, below. The first primary claim refers to the definition of college and career readiness, the following four primary claims relate to the process used to develop performance level descriptors, and the last six are claims associated with setting standards. In addition to these 11 primary claims there are five secondary claims, A.4.6 – A.4.10. These claims are primary claims for A.4, but reference evidence and considerations that will also be relevant to a comprehensive evaluation of A.1. For a complete list of the claims associated with each test characteristics criterion refer to Appendix B.

⁹ The sponsoring agency is the sponsor of the evaluation, and may be the state, a large school district, a private school or some other organization.

TABLE 6. PRIMARY AND SECONDARY CLAIMS ASSOCIATED WITH CRITERIA A.1

A.1 Indicating progress toward college and career readiness: Scores¹⁰ and performance levels on assessments are mapped to determinations of college and career readiness at the high school level and for other grades being on track to college and career readiness by the time of high school graduation.

Primary claims related to the definition of CCR

- A.1.1. College- and career readiness has been clearly defined for operational use.

Primary claims related to the performance level descriptors

- A.1.2. The process for developing performance level descriptors (PLDs) provides for PLDs that accurately represent the expectations defined by the CCR content standards within and across grades.
- A.1.3. Knowledgeable experts were involved in the process of developing and reviewing the PLDs.
- A.1.4. The process used for developing performance level descriptors (PLDs) supports their intended use(s).
- A.1.5. The process for developing performance level descriptors (PLDs) includes an evaluation of alignment of the PLDs to the content of the test questions that differentiate performance at each level, and, as needed, re-writing based on new evidence concerning skills needed for success in college and careers.

Primary claims related to standard setting

- A.1.6. A description and coherent rationale are provided for how the proposed and/or implemented standard setting methodology¹¹ yields valid determinations of progress toward, or attainment of, college and career readiness.
- A.1.7. A coherent rationale accompanies methodological decisions regarding the level of involvement of grade-level educators, higher education, industry, and career technical experts (CTEs) in the standard setting process.
- A.1.8. Appropriate external CCR benchmarks and research studies are/were used in the standard setting process.
- A.1.9. Procedures and rationales for any adjustments made to proposed cut scores after the standard setting meeting are based on a defensible rationale and method.
- A.1.10. Studies *planned or conducted* to evaluate the validity of CCR performance standards over time are appropriate given the inferences they are intended to support.
- A.1.11. The standard setting procedures were followed as specified, and the final cut scores and the results of validity studies have been reviewed by technical experts.

Secondary claims from A.4 related to scaling and equating

Evidence related to the design of the reportable scale and the procedures used to translate student performance to that metric may inform decisions around the appropriateness of standard setting procedures, results, and plans for standards validation (specifically claims A.1.6 and A.1.10). Similarly, accuracy in the equating process is necessary to ensure that cut scores do not drift away from their true/intended value over time.

The sufficiency/quality of the evidence presented in relation to claims A.4.6-A.4.10, therefore, should be taken into consideration when evaluating the claims associated with A.1, and when making a final, holistic determination regarding the strength of evidence presented in support of this criterion.

- A.4.6. The design of the scale accounts for the design of the assessment and the manner in which results are intended to be interpreted and used.
- A.4.7. The procedures used to estimate student performance and translate these estimates to a different scale are transparent, fair, and consistent with the reported meaning of the scale scores.
- A.4.8. Procedures for scoring items or sections that involve human judgment (e.g. performance tasks, essays) support accurate and consistent scoring within and across items, forms, administrations, and sub-groups by minimizing construct-irrelevant score variance within and across scorers.
- A.4.9. Linking and/or equating procedures are clearly specified, comprehensive, and demonstratively appropriate.
- A.4.10. The scaling and linking/equating procedures were followed as specified, and the results have been reviewed and accepted by technical experts.

¹⁰ The claims regarding evidence for relating test scores to college and career readiness indicators as defined for operational use can be found in the validity evaluation section under Criterion A.2.

¹¹ The standard setting *methodology* refers to the specific technique or approach used by panelists to recommend performance standards (a.k.a. cut scores) within the context of the standard setting meeting.

It is through the use of primary and secondary claims that the CEF serves to highlight important relationships among the criteria. Specifying connections across criteria at the claim-level serves to clarify how and why key components are related. We hope this is much more informative than simply saying that two criteria are connected or that the evidence associated with one criterion may be relevant to another. To make this link even more transparent, in addition to listing the secondary claims, the CEF provides a brief summary of why the identified secondary claims are relevant to the review of that criterion, as illustrated in Table 6.

While the claims define *what* must be reviewed to evaluate each of the CCSSO criteria, they do not dictate *how* those materials should be reviewed or the means by which decisions about the quality, appropriateness, and sufficiency of that evidence should be determined. Therefore, for each claim we provide examples of what evidence for high quality assessments should look like and provide comments to inform the evaluation of these examples in different contexts. These elements, represented by the bottom two levels of the framework, are referred to as **sufficiency statements** and **comments**, respectively.

Sufficiency statements describe those features/characteristics we believe should be reflected in a particular type of evidence in order for it to lend useful and adequate support to a given claim.

For many reasons, the manner and degree to which a particular type of evidence exists will differ across programs, as will its importance relative to the evaluation of a given claim. Therefore, while we consider producing evidence to support all claims to be compulsory, sufficiency statements are *not* conjunctive lists of must-haves. Instead, they provide a general indication of what evidence of high-quality should or could look like for a given program. Sufficiency statements are designed to be an illustrative aid for expert judgment rather than the rule. Understanding the goals of the assessment, theory of action, and intended use of assessment results will be critical for evaluating the quality of evidence provided. Those involved in conducting the evaluation will be asked to consider these features (e.g., use, stakes, examinee population, etc.) in addition to the assessment's current phase of development or implementation to make judgments about the degree support provided for each claim and a holistic rating for each of the CCSSO Criteria. To inform this process, comments and examples are provided to highlight how contextual factors may influence one's thinking about the evidence submitted for a given test.

Some of the evidence statements outlined in the sufficiency column may go beyond what is typically developed or collected to support a traditional summative assessment. While it may not be reasonable to expect assessments that are in the operational phase to go back and collect this information solely for the purpose of this review, these examples are included in sufficiency statements to outline the expectations for this type of evidence, when appropriate and relevant, in future test development iterations.

Comments are included as additional notes to aid reviewers in judging the quality of particular evidence within the context of an assessment program.

While the comments provide useful evaluation tips and were developed with input from technical advisors, like sufficiency statements, they are not designed to be comprehensive or compulsory.

The full specification of each criterion is provided in the Criteria Evaluation Framework (CEF). These specifications represent the result of an iterative development process which included the review and provision of feedback from a team of measurement experts with valuable contributions. It is assumed that the CEF will continue to be revised and refined over time as additional feedback is collected—even after it has been applied in operational settings.

Review of Assessment System Outputs

The assumptions outlined in Table 5, Key Design Assumptions and Constraints, had a significant impact on not only the methodology (Part 2), but the manner in which each of the criteria were operationalized. For example, the need for an efficient evaluation process largely influenced the manner in which the review of certain outputs was addressed in the Criteria Evaluation Framework. There are a variety of statistical analyses, psychometric, and administrative procedures conducted to support the development, scoring, reporting and validation of assessments results for their intended uses. Each procedure results in a multitude of outputs (both qualitative and quantitative) which could be evaluated. Within the context of this document “procedures” are distinguished from “outputs” as follows:

Procedures: Descriptive summaries of the participants, analyses, rules and/or criteria employed (or intended) to support or used to implement a given test development or administration activity.¹²

¹² Additionally, much of the evidence provided to support claims about data privacy and test security will be procedural.

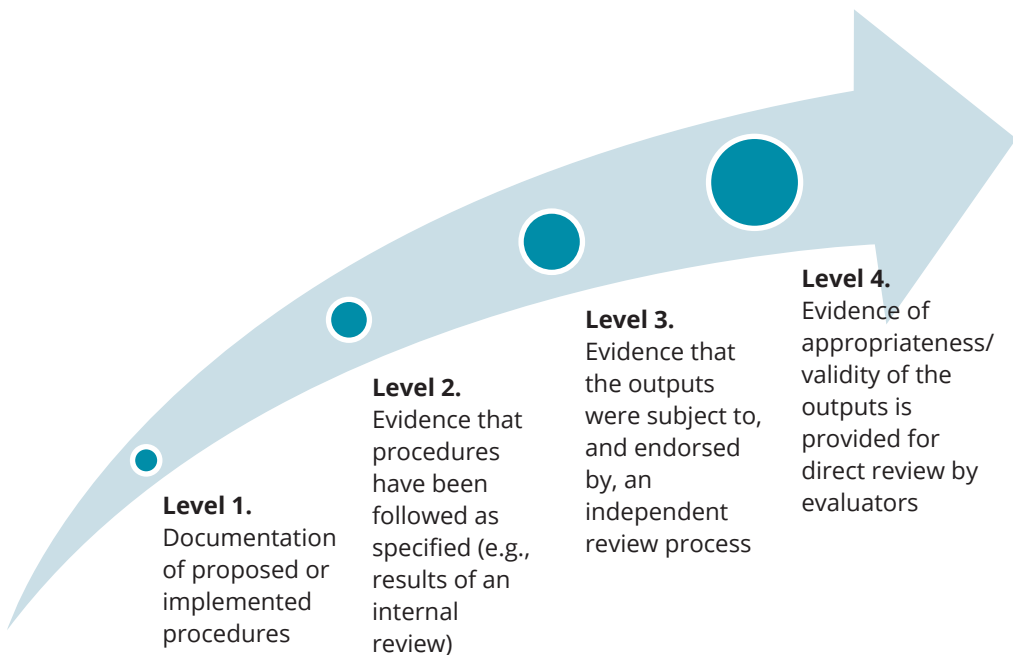
Outputs: The data, information, results, decisions and inferences resulting from the implementation of a given assessment-related activity.¹³

Outputs include reports on, or documentation of: item analyses, equating, test development, performance level descriptors, performance standards, and score reports. In certain situations outputs may also include notes or minutes resulting from an expert review of the procedures and results associated with an activity (e.g., equating, scaling, forensic data analysis), and/or a report penned by an independent technical advisor summarizing that established procedures were conducted with fidelity (e.g., standard setting). Outputs also include results from validation studies (e.g., of equating methods or calibrations), technical audits of key procedures and independent replication studies.

For most assessment programs, the amount of time and detailed contextual knowledge necessary to conduct a comprehensive evaluation of *all* of these elements will far surpass that available to conduct an efficient evaluation. Therefore, to increase the utility of this work, the CEF does not require (or suggest) that evaluators review every output resulting from the implementation of a specified process or analysis. Instead, the Criteria Evaluation Framework may ask reviewers to evaluate the quality of: 1) evidence describing the *procedures* used to generate those outputs and their associated rationales, 2) evidence demonstrating that the *procedures* were conducted in the manner specified and/or 3) evidence demonstrating that qualified technical experts were involved in the review and endorsement of *procedures* and *outputs*. Only in some cases, where it is believed to necessary to support evaluation, will outputs resulting from procedures be explicitly identified as necessary for review.

These types of evidence and the manner of review they afford differ greatly, and can be conceptualized as falling along a continuum ranging from least to most rigorous within the context of an evaluation. A visual representation of this continuum is reflected in Figure 3.

FIGURE 3. PROCEDURE- AND OUTPUT-BASED EVIDENCE CONTINUUM



The far left side of the continuum (i.e., least rigorous) represents instances where procedures, rather than outputs are specified for review. For example, within the CEF, evaluators are asked to review the materials, participants and training documents supporting the PLD development process, but they are not asked to evaluate the quality of the final PLDs. That is not to say that outputs associated with the PLD development process will not be reviewed. If available, panelist survey results, participant demographic summaries, and/or external reports reflecting the integrity of the PLD process and results may be provided as evidence. It is simply that the PLDs themselves will not be targeted for review. As you move along the continuum, the evaluation becomes more output-based. For example, the second point in the continuum reflects the fact that for some activities, evidence must go beyond the specification of processes to include documentation demonstrating an internal review and/or replication of

¹³ It is understood that some outputs (e.g., equating reports) may include descriptions of procedures, but we would still refer to the procedures implemented as separate from the outputs.

key outputs and results occurred. The replication activities are not something the evaluators would be asked to conduct, instead they would be asked to evaluate whether strong evidence was provided that such activities occurred as planned. For more important pieces of evidence, evaluators will be asked to review evidence that processes were followed as planned and that key outputs were reviewed and approved through an independent review process, as represented by the third point in the continuum. This includes procedures such as equating, scaling, and standard setting for which a comprehensive review and approval of all analyses, outputs and decision points is not feasible in the time allotted. Finally, some evidence is important enough—such as samples of reliability coefficients and results of validity studies—that evaluators will be asked to review the outputs themselves, regardless of any other review activities that may have occurred prior to evaluation.

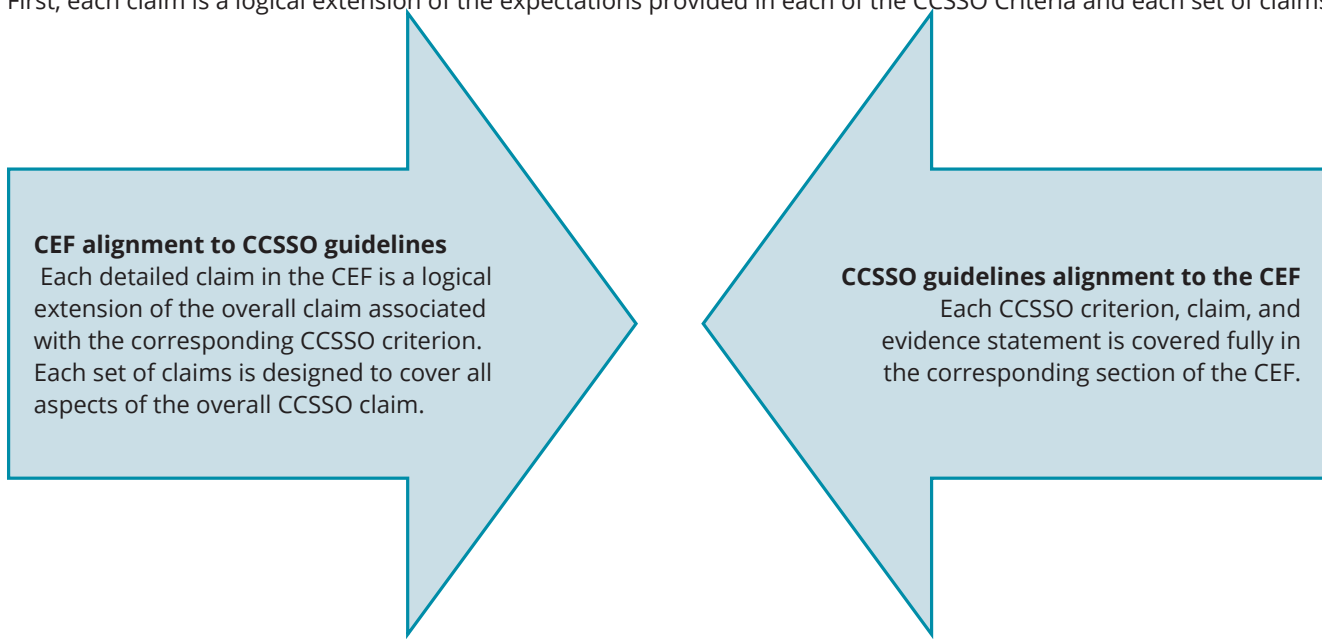
To determine the appropriate level of review to incorporate into the CEF for a given activity, Center for Assessment researchers considered which evidence could and could not be defensibly evaluated within a reasonable timeframe. For example, when it comes to equating, we believe it is reasonable to ask evaluators to consider documentation summarizing the equating procedures and evaluation criteria, as well as evidence showing that the equating process was implemented as expected and reviewed by a panel of third party technical experts. However, it is not reasonable to ask evaluators to review the detailed equating outputs associated with all grades and subjects and make an overall evaluation of the quality of these results for a given administration. Doing so requires data, information and insight that the evaluators will likely not have access to and, in many cases, may not be equipped to evaluate. In addition, it would take a significant amount of time and effort on the part of the reviewers as the results associated with one grade and content area are not generalizable, making sampling to support evaluation difficult to defend. In lieu of this, the evaluators are asked to review evidence demonstrating that 1) specified quality control procedures were implemented (e.g., replication of results) and 2) a comprehensive review of the equating results was conducted by a set of qualified technical experts (i.e., Level 3 on the process and outcome-based evidence continuum). Such evidence would include not only documentation that a technical review occurred, but also the credentials of those involved, the type of data/materials reviewed, the nature of the feedback obtained (e.g., the underlying results regarding the strength or weakness of the equating) and a summary of how that feedback was addressed or acted upon.

Relationship between the Framework and Original CCSSO Documentation

Alignment between the original CCSSO Criteria and the Criteria Evaluation Framework is two-fold, as demonstrated in the figure below.

FIGURE 4. ALIGNMENT BETWEEN CCSSO CRITERIA DOCUMENT AND CEF

First, each claim is a logical extension of the expectations provided in each of the CCSSO Criteria and each set of claims, as a



whole, is designed to comprehensively and directly support evaluation of a given criterion statement. Secondly, each criterion, example, and consideration presented in the CCSSO Criteria document is included somewhere within the Criteria Evaluation Framework. A crosswalk was designed as a quick-reference for identifying where each idea from the CCSSO Criteria document is represented in the CEF is provided in Appendix A. An example of this crosswalk is included in Table 7.

TABLE 7. EXCERPT OF CROSSWALK DOCUMENT

<p>A.7 Meeting all requirements for data privacy and ownership: All assessments must meet federal and state requirements for student privacy, and all data is owned exclusively by the state.</p>	<p>All six claims associated with Criterion A.7 are logical extensions of the language expressed in the original CCSSO document and are designed to specify a comprehensive set of evidence that supports evaluation. The specific claims listed in column 3 below indicate where each of the evidence statements presented in the CCSSO documentation can be located within the Criteria Evaluation Framework.</p>	
	<ul style="list-style-type: none"> • An assurance is provided of student privacy protection, reflecting compliance with all applicable federal and state laws and requirements. 	A.7.1
	<ul style="list-style-type: none"> • An assurance is provided of state ownership of all data, reflecting knowledge of state laws and requirements. 	A.7.4
	<ul style="list-style-type: none"> • An assurance is provided that state will receive all underlying data, in a timely and useable fashion, so it can do further analysis as desired, including, for example, achievement, verification, forensic, and security analyses. 	A.7.5
	<ul style="list-style-type: none"> • A description is provided for how data will be managed securely, including, for example, as data is transferred between vendors and the state. 	A.7.3, A.7.6

All of the text in the table above comes directly from the CCSSO Criteria document. The rightmost column contains an index of the claim in the CEF that addresses each of those ideas. Close review of the crosswalk table reveals that: 1) all of the CCSSO examples of evidence appear somewhere within Criteria Evaluation Framework, and 2) many claims in the CEF represent extensions of the original CCSSO Criteria (i.e., are not explicitly represented in the original document). While the CEF development process initiated with the CCSSO examples of evidence, for reasons previously discussed, additional detail was added to support the development of a comprehensive, coherent evaluation process. We believe this is a rational approach to building upon CCSSO's document in a way that conveys, in broad strokes, the criteria by which assessment quality should be judged.

PART 2: GUIDE FOR IMPLEMENTATION

Summary of the Test Characteristics Evaluation Methodology

Throughout the discussion which follows, we consistently refer to four different entities, each of which has a clearly defined role in the overall evaluation process. For clarity, each group and its expected role is outlined in Table 8.

TABLE 8. GROUPS INVOLVED IN THE TEST CHARACTERISTIC CRITERIA EVALUATION METHODOLOGY

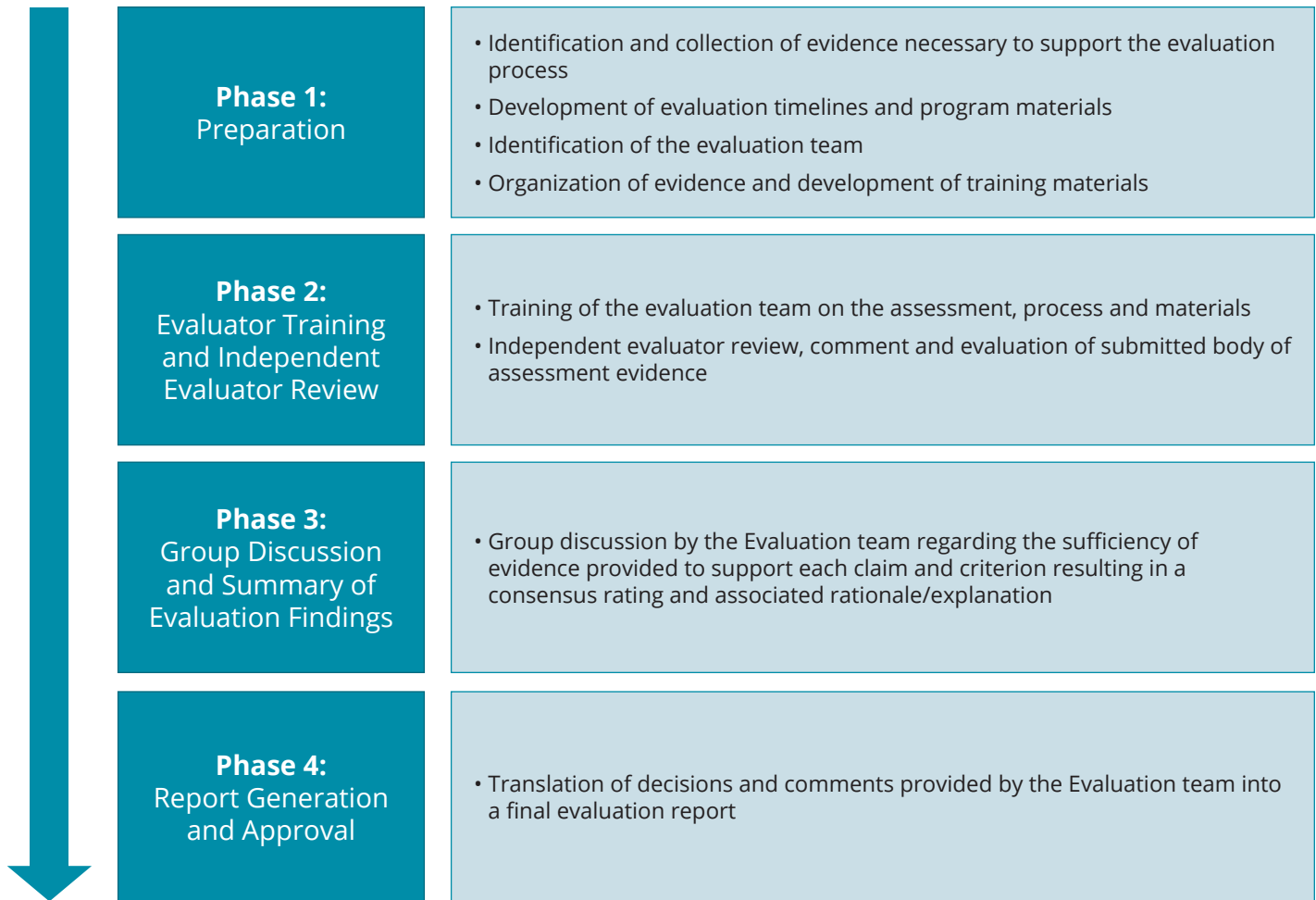
GROUP	ROLE
<p>Requester: The group or organization requesting the evaluation.</p>	<ul style="list-style-type: none"> - Identify the test(s)/assessment programs to be evaluated, the purpose of evaluation and the intended user of the evaluation¹⁴ results - Outline the manner in which evaluation results are (or are intended to be) used
<p>Implementer (a.k.a. Implementation Team): The group or organization responsible for conducting the evaluation in adherence to this methodology.</p>	<ul style="list-style-type: none"> - Coordinate and schedule the evaluation activities - Identify and contact appropriate technical experts to act as evaluators - Work with the Provider to identify all vendors/parties that will need to supply evidence to support evaluation - Obtain evidence from the Provider and ensure it is organized in the manner specified - Make sure the Provider understands the rules related to the provision of evidence (e.g., meeting all requirements the first time around) - Create secure repository for materials and provide expert reviewers with access - Provide training to experts on the Independent Evaluation process and the format/ location of provided materials - Facilitate the Evaluation Team meeting - Generate final evaluation report based on Evaluator decisions and feedback
<p>Provider: The primary group/ organization responsible for providing information to the Implementation Team. The Provider may need to contact multiple groups (state agency, companies, university, organizations, technical teams) - those who are, or will be, responsible for the design, development, scoring, reporting, security and administration of the test – to acquire all of the materials necessary for evaluation.</p>	<ul style="list-style-type: none"> - Identify, gather and organize appropriate evidence to inform the evaluation process - Support the development of a general overview/summary of the assessment that outlines, phase of development or implementation, history of the assessment, purpose for which it was developed, etc. - Address questions posed by Implementers and requests for clarification (as necessary) - Provide brief comments in response to the final report (if desired)
<p>Evaluators (a.k.a., Evaluation Team): The technical experts charged with reviewing and evaluating the submitted evidence.</p>	<ul style="list-style-type: none"> - Review the evidence provided to support evaluation - Make comments and ratings about the adequacy of that evidence relative to particular claims and the CCSSO criteria - Comment on and approve the final report

The evaluation methodology will be discussed in four phases, as summarized in Figure 5. In the following section of this report, each of these phases is discussed in detail providing guidance to those participating in the evaluation. It is important to note that the methodology outlined below is written to address a situation where one assessment program is selected for evaluation. Methodological considerations specific to the evaluation of multiple assessment programs simultaneously (for comparative purposes), and assessments developed to serve multiple states are addressed later in this document.¹⁵

¹⁴ The purpose of the evaluation differs from the “goals” specified in Table 7 in that the stated purpose will be specific to the current or proposed evaluation. For example, “the purpose of the proposed evaluation is to aid in the assessment contract renewal decision for academic year 2016-2017 regarding the current alternate testing program for students with severe cognitive disabilities in State X

¹⁵ See “Considerations related to the Implementation of the Methodology for Multiple Assessment Simultaneously”

FIGURE 5. PHASES OF THE TEST CHARACTERISTICS EVALUATION METHODOLOGY



Phase 1: Preparation

There are three major preparation activities that must occur before an assessment evaluation can be initiated: the evidence necessary to support review must be acquired, an appropriate Evaluation Team must be identified, and the review materials must be assembled in a secure repository. Each of these activities is discussed in the sections below.

Submission and Organization of Evidence Necessary to Support Review

From the Criteria Evaluation Framework, it is clear that a great deal of evidence is necessary to support an evaluation of test quality relative to the CCSSO Criteria. For a given assessment this evidence may take several forms (e.g., technical reports, research studies, meeting minutes, etc.) and be collected from a variety of different sources (e.g., test development, scoring and administration vendors, State Department staff, etc.). Given the scope of this activity, as soon as an evaluation is requested, the Implementation Team's primary goal should be to work with the Provider to identify these sources and send out formal requests for evidence. Locating the evidence may be difficult and will likely require the joint efforts of the Implementation Team and the Provider. Clear communication about exactly what evidence is being requested will be necessary as any one piece of evidence may take a variety of forms across different vendors.

Given the detailed nature of the evidence required, the request for information should make it clear that those responsible for identifying and compiling evidence must have a comprehensive understanding of the assessment program and the technical requirements underlying each of the CCSSO criteria (i.e., support staff likely do not fit this requirement). Appendices B and C are included at the end of this document to help inform this process. Appendix B contains the complete structure of the CCSSO Criteria associated with test characteristics and their supporting claims, and the Sample Evidence document in Appendix C provides *examples* of different types of evidence that may be submitted to support the evaluation of each criterion.¹⁶ A version of this document should be provided in conjunction with any request.

¹⁶ The sample document in Appendix C provides examples that apply specifically to pre-existing assessments.

For reasons previously discussed, it is understood that the type of materials provided for evaluation will vary across tests due to contextual factors such as the assessment's phase of development or implementation. Different evidence will be expected from an assessment that is newly proposed (for example, in response to an RFP) compared to an assessment that is in the initial phases of development or one that has had one or two years of full implementation. The Criteria Evaluation Framework articulates, primarily, what one would expect to see from an assessment that has already been implemented. When appropriate, comments indicate how expectations are influenced by phase of development or implementation, but in most cases it will be up to the evaluator to determine whether the range of evidence provided aligns with his/her expectations given the assessment's age and design. Similarly, those charged with identifying and assembling evidence for review (i.e., within the Provider organization) are responsible for determining what pieces of evidence are necessary, and would be expected, to support the evaluation of each claim. In doing so, they should strive to *identify the minimum amount of documentation necessary to allow for qualified technical experts to make consistent determinations regarding the extent to which a given claim has been met*. In other words, the evidence provided should be detailed and comprehensive, but it must not be a "data dump" that requires the evaluators to sift through piles of marginally relevant materials and determine what is important.

Even though the primary responsibility of determining what materials are necessary and how they should be presented falls to the Provider, the Implementation Team is ultimately responsible for ensuring an efficient, effective review. The Implementation Team will verify that all evidence obtained from the Providers is clear, appropriate and well organized. For this reason, the Implementation Team must have at least a general working knowledge of the type and scope of materials necessary to support each claim so they can verify the appropriateness and alignment of the materials provided.

To support the assembly and presentation of evidence, two sample templates are provided in Appendix D. The first template is the Evidence Log (E-log). The E-log should be used to assign each piece of evidence a unique document number and briefly describe the information the evidence supplies. The second template is the Prioritized Evidence List. For each criterion, the Prioritized Evidence List outlines the evidence assembled for each claim *in order of importance or relevance*. Further instructions and specifications for the type of data and information that should be included on each of these templates are provided in Appendix D.

It is important to note that there are special considerations related to the collection and organization of evidence for assessments developed to serve multiple states (e.g., ACT, SAT, PARCC, SBAC, etc..) This and other issues related to implementation of the methodology with multi-state assessments are addressed in the section titled "Considerations Related to the Implementation of the Methodology for Assessments Developed for use in Multiple States."

Identification of the Evaluation Team

The evidence necessary to support the evaluation of a given assessment depends on the purpose(s) of the assessment and the score-based interpretations necessary to use test results as intended. While the interaction between these factors and the evidence expected/required to support each evaluation is partially reflected in the Criteria Evaluation Framework, the unique interplay of all possible influences can never be fully addressed. For this reason, those selected to conduct the evaluation must have not only a deep understanding of applied psychometric issues, but also how they interact with contextual factors to influence decisions regarding the quality, relevance and sufficiency of evidence. In addition, because there are important relationships among criteria, in most cases evaluators will be reviewing the evidence submitted for all of the test characteristics criteria (as opposed to a subset), therefore, broad expertise is required.

It is the Implementation Team's role to identify an appropriate set of technical experts to support evaluation. This activity should begin as soon as possible after the request for evaluation so training and review activities can begin in a timely manner after the assembly of evaluation materials. Some guidelines supporting the selection of Evaluators are provided in Table 9.

TABLE 9. CONSIDERATIONS RELATED TO THE SELECTION OF EXPERT EVALUATORS

1. Target number of Evaluators:

The appropriate number of evaluators will depend on the context surrounding the evaluation and the range of expertise represented in the evaluator panel, as previously discussed. However there should be enough evaluators involved in the process to allow for each criterion to be discussed and evaluated by at least two experts. In addition, in no case should the minimum number of evaluators fall below four for a given assessment.

2. Qualifications for Individual Evaluators:

- Proven *applied* and *technical* expertise in the field of educational measurement and assessment. Evaluators need to be conversant with operational issues as well as technical issues.
- Appropriately independent from the assessment to be evaluated: How “appropriately independent” is operationalized may vary across tests and settings, but in no case should an evaluator be associated with the test under review in a manner that would make him/her feel inclined or obligated to defend the evidence provided for personal or professional reasons. The Lead Psychometrician associated a customized state assessment, for example, would not be an appropriate evaluator because they have a vested interest in the test. Similarly, a TAC member associated with the state test under review may have provided a level of guidance/support which makes it difficult to be impartial. A TAC member associated with a state using a consortia-developed assessment, however, may be a completely appropriate selection. The Implementation Team will need to determine which experts should/should not be considered given their unique circumstances and recruit accordingly.

3. Composition of the Evaluation Team:

- Although each evaluator should meet the qualifications outlined above, any focal areas of expertise (e.g., assessment of students with disabilities, Equating/Scaling; Value-Added Models; Validation, etc.) should be distributed across evaluators and take into account such factors as the intended student population and manner in which assessment results will be used.

The Implementation Team will need to determine, through vitae review and conversations with potential evaluators, which experts can support the review of multiple criteria.¹⁷ If, for some reason, each evaluator does not have sufficient expertise to fully evaluate all 9 criteria, the criteria may be divided into subsets with each expert reviewing only those criteria aligned to his/her area of expertise. For example, given the unique background knowledge necessary to evaluate the claims and evidence associated with criterion A.5, it may be necessary to identify a separate pair of evaluators for this criterion. Throughout the remainder of this document we refer to this as a *matrix-based evaluation design*.¹⁸

When a matrix-based evaluation design is applied, it is our strong recommendation that each criterion be reviewed by at least two experts so that discussion about the quality and appropriateness of evidence can occur. When the criteria are divided, the primary goal is to ensure that the team as a whole, rather than each individual, represents the range of skills and expertise needed to conduct a comprehensive evaluation. In those situations where multiple assessments are being evaluated concurrently a matrix-based design may be considered for logistical reasons. In this case, evaluators might be assigned a common subset of the test characteristics criteria to review across the assessments to be evaluated, rather than the full body of evidence submitted for each test. Procedures supporting the review of multiple assessments are discussed later in this document.

When the Implementation Team contacts potential evaluators it is important that they provide them with all of the information necessary to make an informed decision about participation. Such information would include: the assessment to be evaluated, a detailed timeline for implementation including any scheduled meeting dates, an estimate of the amount of time and the activities associated with participation, information about disclosure (e.g., will their name be publically associated with the evaluation), and required qualifications.

¹⁷ During the test case of the Assessment Quality process evaluators were quick to note if they did not feel as if they had the expertise to offer a rating for a particular criteria. This only happened in one case, and it was for Criterion A5.

¹⁸ Please note that this is not the preferred or recommended design, as we believe it is optimal to have each evaluator review all submitted evidence. That being said, we understand that this may not always be possible.

Secure Assembly of Evaluation Materials

Some of the evidence provided to support assessment review may be considered confidential or proprietary (e.g., test items, forms, draft procedural documentation, security protocols, etc.) by the Provider. To ensure confidentiality and security are maintained, all procedures related to the storage, delivery, and removal of submitted evidence should be clearly documented by the Implementation Team and provided in concert with the initial request for information. It is the Implementation Team's job to make sure the materials submitted to support evaluation are kept secure and access is provided only to those who need it. It is the responsibility of each evaluator to maintain the confidentiality of materials provided to them. Furthermore, once the evaluation is concluded, procedures should be in place to ensure all materials are securely purged or provided to the Requester in a manner consistent with that agreed upon by each participating organization.

The specific procedures and tools used will be developed by each Implementation Team, but should include such things as:

- Password protection for access to any files to support review.
- Signing of non-disclosure agreements and avoidance of conflict of interest by all individuals granted access to evaluation materials.
- The use of file encryption procedures and/or a secure drive for uploading and sharing materials.
- In the case of paper-based materials, a log for managing the location of documents including dates of distribution to the Evaluators and return to the Implementers for destruction or delivery back to the Provider.

To help improve the efficiency of the evaluation process, the Implementer is also responsible for checking that all submitted materials are organized and indexed in a way that facilitates accessibility to the Evaluation Team. This includes reviewing the Evidence Log to make sure each document, or set of documents, has an associated document number, verifying that the Prioritized Evidence List has been completed as expected, and each piece of provided evidence appears at least once. Any questions related to the organization of evidence or its purpose should be identified and handled by the Implementation Team prior to release to the Evaluation Team.

Phase 2: Evaluator Training and Independent Evaluator Review

The primary activities associated with Phase 2 of the methodology include 1) evaluator training on the assessment, evaluation materials and methodology and 2) independent evaluator review and evaluation of all submitted evidence. While example procedures for conducting Phase 2 of the methodology are provided below, this represents one of many possible different ways in which these activities may occur. Specifics related to the manner in which panelists are convened (e.g., webinar vs. face-to-face) and ratings are collected (e.g., on paper or electronically), for example, may vary depending on the scope of the evaluation effort and timeline for implementation. Users of this methodology should consider the procedures outlined below as one example of how Phase 2 may be conducted and modify the details as necessary to meet their needs and circumstances.

Evaluation Team Training

Prior to the scheduled window for independent evaluation, the Implementation Team should conduct training for the Evaluators that serves to:

- Introduce the assessment to be reviewed.
- Provide an overview of the phases of the evaluation
- Train evaluators for the independent review process

To make the training as efficient and useful as possible, the Executive Summary and the Criteria Evaluation Framework should be provided to the Evaluation Team for required review prior to the meeting.

Introduce the Assessment

The assessment overview should be brief, but comprehensive, focusing on background and contextual information relevant to the evaluation process. This would include such things as the age of the assessment, the target population, the standards the assessment was developed to address, the extent to which the standards have been taught/assessed in the state, the purpose of the assessment and intended uses of assessment results, and any other *relevant* factors that should be considered or taken into account when conducting the evaluation.¹⁹ The Implementation Team should work with the Requester and/or the Provider to identify and articulate this information. Often, it may make sense to have one of these parties develop and deliver the presentation. However, the overview is *not* intended to be technical, so the content should be appropriate for presentation by the Implementation Team. The Provider may be part of this training and/or standing by to answer any questions; however, to

¹⁹ Note: This introduction may need to be more comprehensive for evaluations in which all evaluators are not reviewing the entire body of submitted evidence (e.g., matrix evaluation design). See section titled, "Considerations Related to the Implementation of the Methodology for Multiple Assessments Simultaneously."

avoid a barrage of detailed technical questions, Evaluators should be informed at the onset that the overview is intended to provide only the assessment context necessary for review. If the focus of evaluation is an assessment proposal (e.g., in response to an RFP), rather than an existing operational form, the introduction should serve to outline key components of the proposed assessment design and associated timeline for implementation.

In developing the assessment overview, it may be difficult, especially for the Implementation Team, to discriminate relevant contextual information from that which may be inappropriate (i.e., could be misused) or distracting. While it is not possible to anticipate how each piece of contextual information will be used to support decision making, one may discriminate information that is/is not appropriate for the overview by posing the following questions:

- Is this something that the evaluators need to know to understand and evaluate the nature and scope of the evidence to be reviewed? If yes, include it.
- Is the information non-technical in nature or is “technical” information presented at a level of detail appropriate for an overview? If yes, include it.
- Does the information represent something that should be considered when evaluating the adequacy, sufficiency and quality of processes and outputs AND may not be provided within the context of those documents? If yes, include it.
- Could this information influence or bias an evaluator’s review of submitted evidence? If yes, leave it out.
- Does the information reflect an influential, pre-existing opinion or position related to the quality of the assessment under review? If yes, leave it out.

A few examples that fall in the “Appropriate” vs. “Not Appropriate” categories are provided in Table 10.

TABLE 10. CONTEXTUAL ELEMENTS TO INCLUDE IN THE OVERVIEW PRESENTATION

APPROPRIATE	NOT APPROPRIATE
<ul style="list-style-type: none"> • Who requested the evaluation and for what purpose. • The intended purpose of the assessment and use of assessment results²⁰ • Historical timeline of events related to the development and roll-out of the assessment and its associated content standards; standard setting, etc. • Grades and subject areas tested • Intended test taking population and size • Time and frequency of administration (i.e., spring, fall, etc.) • Number and type of accommodated forms available for use • Important legislative requirements, such as those linking test performance to graduation or retention. • Phase-in or transition plans that influence assessment design or outcomes • Key terminology and acronyms • Mode of administration 	<ul style="list-style-type: none"> • Whether there are resource issues (financial, staffing) • Technology constraints • Legislative constraints • Outside opinions related to the quality of the assessment or its utility

Some might argue that issues related to availability of resources are important to consider when evaluating the sufficiency of evidence provided. For example, if a test vendor is not supplied with the funds necessary to collect or create evidence of the type/scope proposed by the Criteria Evaluation Framework, it may be viewed as unfair to rate a claim or criteria as not meeting expectations since non-compliance is out of their control. However, the goal of this methodology is to provide feedback about the quality of evidence available to support each of the CCSO criteria, therefore, the focus of evaluation should be on identifying gaps in evidence regardless of why those gaps may have occurred.

Ratings and comments can only be made in light of evidence reviewed. An assessment program may state that they used “psychometrically sound and defensible procedures for establishing college-readiness benchmarks”, but without evidence to support this, it is nothing more than an unsupported claim. If there is no evidence submitted to support the evaluation of a

²⁰ For existing assessments it is especially important to understand the purpose/uses the assessment was designed to support and how (or if) they differ from the manner in which the test is intended to be used. This methodology is developed to evaluate summative assessments aligned to CCR standards. If an assessment being evaluated was not initially designed for that purpose this must be understood.

given claim or criterion, we strongly encourage Providers to supply a rationale for the omission. Ultimately, the Evaluation Team will determine how, and or if, this rationale should influence their claim and criterion-level ratings.

By the end of the overview the Evaluation Team should have a clear understanding of the assessment's phase of development and/or implementation. To ensure this understanding is both shared and considered throughout the evaluation process, the Implementation Team should classify the assessment in terms of where it exists on this continuum.²¹ The classification should be agreed upon by all parties (Provider and Evaluators), as it will also be highlighted in the final report to support the interpretation of ratings. A proposed classification scheme is outlined in Figure 6 (which appears later in this document); however the Implementation Team may decide that a different set of categories or level of specificity should be used given the goals of evaluation.

Discuss the Different Phases of the Evaluation Methodology

Once the assessment has been introduced, the Implementers should provide a brief overview of the evaluation methodology and the materials that will be used to support it. This is a good opportunity for the Implementation Team to discuss the “nuts and bolts” of the evaluation, including the timeline, responsibilities and report writing. Evaluators should be informed that details and training specific to the Independent Evaluator Review will follow this general overview.

Since much of the work occurring in the Preparation phase is explicitly tied to the content and format of the Criteria Evaluation Framework (e.g., identifying contributors; collection and organization of evidence) an overview of this document should occur early in the training process. Specifically the Implementation Team should:

- Briefly describe process used to develop the Criteria Evaluation Framework (as outlined in this document).
- Define the elements of the CEF (e.g., criteria, claims, sufficiency statements, etc.) and the specific information/guidance provided by each.
- Highlight the relationships across criteria, in terms of the representation of secondary claims.
- Highlight the fact that the Criteria Evaluation Framework is intended to be used as a guide to support evaluation relative to the CCSSO Criteria (i.e., that it was developed to be comprehensive, but not exhaustive).
- Discuss the range of outputs the team will be responsible for reviewing and how/why it varies across criteria.

After this discussion, the Evidence Log and the Prioritized Evidence List can be distributed and discussed. The Implementation Team should describe the content, format and purpose of each these documents, highlighting features relevant to the evaluation process (e.g., the same piece of evidence may be associated with multiple claims; for each claim the evidence list has been prioritized by importance). If the Provider encountered issues in the identification or collection of evidence, specifically those which might influence the evaluation process, this should also be discussed (e.g., if there were claims or criteria for which evidence was not provided).

After the Preparation discussion, each of the remaining phases should be broadly introduced, as summarized in Figure 5, with an assurance that the procedures, materials and outputs associated with each phase will be fully addressed in training.

Training on the Independent Evaluator Review Process

After introducing the components of the overall evaluation methodology the Implementation Team should describe, in detail, the activities and expectations underlying the Independent Evaluator Review. This training has 4 important components:

- Outline the purpose of the Independent Evaluator Review
- Describe procedures for accessing evidence and security protocols
- Describe the steps in the Independent Evaluator Review
- Calibrate evaluators on the claim-level rating process and scale

Purpose of the Independent Evaluator Review

Evaluators should be informed that the primary purpose of Independent Evaluator Review is to provide each evaluator with unstructured time to consider submitted evidence relative to the Criteria Evaluation Framework and develop an informal argument regarding the extent to which each of the claims and criteria statements are supported. By the end of the independent review process it is intended that each evaluator will have established preliminary claim and criterion-level ratings and associated rationales.

While the goal is to have all evaluators reviewing the full range of evidence submitted for each of the CCSSO Criteria, as previously discussed, in some cases this will not be possible resulting in the use of a matrix-based evaluation design. In these

²¹ If tests associated with different grades or subject areas are at different phases of development this should be clearly noted.

situations evaluators may be assigned pairs or sets of criteria to review based on their unique expertise. When working under this design, evaluators should be notified of their assigned criteria at the beginning of the training process so that they can identify questions and concerns with this knowledge in mind.

While a deadline for completion will be defined, the Implementation Team can decide how evaluators work within that timeline. The process may be designed so that the evaluators have complete control over their processes, or the Implementation Team may want to structure a set of intermediate goals with interim check-ins to identify and remedy any issues in the process that may arise during independent review. Evaluators should direct their questions about particular pieces of evidence to the Implementation Team. The Implementation Team will work the Provider, as appropriate, to obtain answers to questions about the supplied evidence.

During training, evaluators should be reminded that the overall purpose of the evaluation is to provide intended users with information about the extent to which *the evidence provided* supports the specified CCSSO quality criteria. Therefore, prior knowledge about, or experience with, the assessment should not be considered when establishing the rating associated with a given claim or criterion. If, for example, one is aware of procedures, practices or research relative to a particular assessment that are not documented or well-articulated within the supplied evidence, comments about such information may be included in the evaluator's review, but should not influence the overall sufficiency rating. If there is important evidence that should have been provided, and is known to exist, but does not appear in the submitted body of evidence, an evaluator may ask the Implementation Team to request this documentation from the Provider so the entire Evaluation Team can review. The Evaluation Team cannot; however, randomly ask for information or evidence that they think (or would expect) an assessment program to have, but was not submitted in the original evidence log. That is, the Provider will not be given the opportunity to generate new evidence to support the evaluation once the original submission of evidence is made.

Accessing the Body of Evidence

After discussing the purpose of the independent review, the Implementation Team should distribute any security codes/ passwords required to access the evidence listed in the Evidence Log and demonstrate any login procedures.²² Although non-disclosure agreements will have already been signed, the importance of maintaining the security of all materials and the need for secure destruction of anything printed should be repeatedly emphasized.

Although there will not be time to walk through every piece of evidence submitted for review, the Implementation Team should select one or two examples to illustrate how the structure and format of the repository align with the information provided in the Evidence Log and Prioritized Evidence List. If there are pieces of evidence that do not exist electronically, these should be indexed in the Evidence Log and provided to the Evaluation Team prior to evaluator training. On occasion, evaluators may request that all of the materials submitted for review be provided in a hard-copy format. Given the large number of documents involved, it is up to the Implementation Team to determine whether this is a reasonable request. Similarly, there may be documents identified during the assembly of materials that the Implementation Team considers important enough to print for *all* evaluators, such as detailed background or technical documents that align to multiple criteria and claims (e.g., Theory of Action, Technical Reports, etc.). In this case the Implementation Team should develop a binder, tabbed by document number, that contains these documents and indicate this fact on the Evidence Log.

The Implementation Team should stress that only those materials provided in the Evidence Log should be used to support evaluation. Evaluators may *not* look to outside sources to find additional information or evidence (e.g., newspapers, blogs or websites) to supplement their review. If there are questions about how a piece of evidence should be interpreted, or potential supporting evidence that is referenced but not provided, the Implementation Team should be notified. It is the Implementation Team's responsibility to provide a designated contact for the Evaluation Team during the course of the independent review phase.

Process for Implementing the Independent Review

Although each evaluator will be able to work through the submitted evidence in a manner that works best for him/her, the Implementation Team is responsible for articulating a coherent evaluation strategy that ensures the expectations for evaluation are consistent and understood across participants.

An exemplar evaluation process is outlined below. This process should be discussed with the Evaluation Team during training, so that expected activities and associated materials are clearly understood. Throughout this section and the remainder of the methodology, there are sub-sections indicated as "Training Notes". The notes include specific recommendations for organizing

²² All evaluators should have signed and submitted non-disclosure agreements prior to the webinar meeting. This could be part of the contract agreement materials.

materials, training evaluators and facilitating the large group discussion given previous experience and feedback provided by evaluators during the test-case of the methodology. We strongly recommend that Implementers adhere to these recommendations when developing training protocols and materials.

Proposed Process for Conducting Independent Evaluation of Assessment Evidence

Summary of Six-Step Process

1. Read through entire CEF
2. Review the Evidence Log
3. Review evidence submitted for Criterion A.1 to make informal ratings at the claim-level
4. Review the evidence for the remaining Criteria
5. Based on all evidence reviewed, make preliminary holistic ratings at the criteria-level
6. Submit forms with ratings and comments to Implementation Team

Step 1: Prior to initiating your evaluation, read through the entire Criteria Evaluation Framework. The primary purpose of this review is to clarify the expectations associated with each criterion and the type/range of evidence that might/should be provided to support evaluation.

Step 2: Review the Evidence Log to get a feel for the range of evidence submitted for review.

Consider the evidence submitted for each of the nine Test Characteristics criteria, as documented in the Prioritized Evidence List. Identify any core or primary pieces of evidence (i.e., that which informs multiple criteria) that should be reviewed prior to conducting your criterion-level evaluation.

Step 3: Review evidence submitted for Criterion A.1 to make informal ratings at the claim-level.

Working claim-by-claim, review the evidence submitted for criterion A.1. Use the Independent Evaluator Rating Form (see Appendix E) to record key comments and concerns and make preliminary claim-level ratings. Specifically, evaluators should do the following for A.1.1:

- a) In light of the body of evidence reviewed for claim A.1.1, make an overall, holistic determination about the extent to which the evidence provided meets expectations given the sufficiency statements and comments associated with that claim in the CEF, and any contextual factors that would influence the type and quality of evidence available for review.
- b) Circle a preliminary rating for A.1.1 on the rating form (e.g., Does Not Meet, Partially Meets, or Meets) and provide a written rationale for that rating. If the evidence submitted goes against or violates the expectations defined for A.1.1 in any way, this should be clearly indicated in the comment section so it can be discussed with the Evaluation Team.



TRAINING NOTES:

During training the Implementation Team should make it clear that:

- *The CEF is intended to guide evaluation but it will not be exhaustive. Throughout the review process it is important for evaluators to take notes or make comments when there are additional factors considered in evaluating evidence relative to a particular claim or criteria.*
- *We understand that the CEF could have been specified in multiple ways and that you may not agree with all aspects of this specification. For the purpose of this evaluation, please try to adhere to the CEF as much as possible. If there are major structural issues that influence your evaluation of a particular claim or criterion, please flag this for group discussion.*



TRAINING NOTES:

- *If desired, the Implementation Team or Provider may recommend particular documents for initial review rather than leaving this up to each Evaluator (e.g., documents that summarize the Theory of Action, history, purpose/goals/uses and/or design of the assessment.)*
- *If a matrix evaluation design is in place (See Figure 7 for an example) such that different evaluators are assigned to review different sets of criteria, core documents that should be reviewed by all evaluators should be clearly highlighted on the Evidence Log and called-out by the Implementation Team during training.*
- *Documents or pieces of evidence that are associated with multiple criteria should be highlighted in the Prioritized Evidence List and called out during training.*

While claim-level determinations will depend on expert judgment regarding the appropriateness of the evidence provided; to support consistency in the rating process across evaluators an operational definition of each rating category is provided in Table 11.

TABLE 11. CLAIM-LEVEL RATING DESCRIPTORS

RATING	
Does Not Meet	There is either no evidence to support the given claim, or, the evidence provided does not lend support to the claim statement as written. The evidence provided may not be relevant, or is extremely insufficient given the assessment's current phase of development or implementation.
Partially Meets	The evidence reviewed provides some support for the claim, and while the claim is partially supported, the evaluator has reservations about endorsing the claim fully. The evidence provided may be incomplete, or not as comprehensive as expected given the assessment's current phase of development or implementation.*
Meets	The evidence reviewed provides clear and unequivocal support for the claim as written. The evidence provided is comprehensive and appropriate given the assessment's current phase of development or implementation.*

*Note: Not everything within the sufficiency statements need be included in evidence - sufficiency statements only describe what high quality evidence could look like.

c) Repeat steps a and b for Claims A.1.2-A.1.11

Note that the initial review of evidence associated with *secondary claims* should be reserved for consideration within the associated primary criterion. For example, there are 5 secondary claims associated with criterion A.1. These claims are primary for criterion A.4, and therefore should be reviewed for the first time when considering A.4.



TRAINING NOTES:

1. *When discussing Step 3 of this process, the Implementation Team should pause to walk through the Independent Evaluation Rating Form (see Appendix E for sample). This document, provides the evaluators with a common template by which to think through and evaluate the evidence provided for a given criterion and set of claims.*
 - a. *Evaluators should be instructed to complete the rating form for each claim throughout the independent evaluation process with the understanding that large group discussion will occur at this level.*
 - b. *Evaluators should be reminded that the claim-level ratings should be based upon the expectations defined within the Criterion Evaluation Framework. They should not be based on past experience or normative expectations defining what is "good".*
2. *If a matrix evaluation design is being used a slightly different process may be considered. For example, evaluators may be instructed to review the evidence associated with any secondary claims in conjunction with the primary claims as they may not have an opportunity to do so at a later time.*
3. *The presentation of Figure 5 should be used as a jumping off point to highlight and discuss the importance of phase-of-development in evaluating the evidence provided for a given claim. To facilitate this discussion, the Implementation Team should identify 1 or 2 claims and then ask the Evaluation Team to discuss how (or if) their expectations about the type and amount of evidence necessary to support that claim might differ for assessments at different phases of development , or for an existing/off-the-shelf test.*

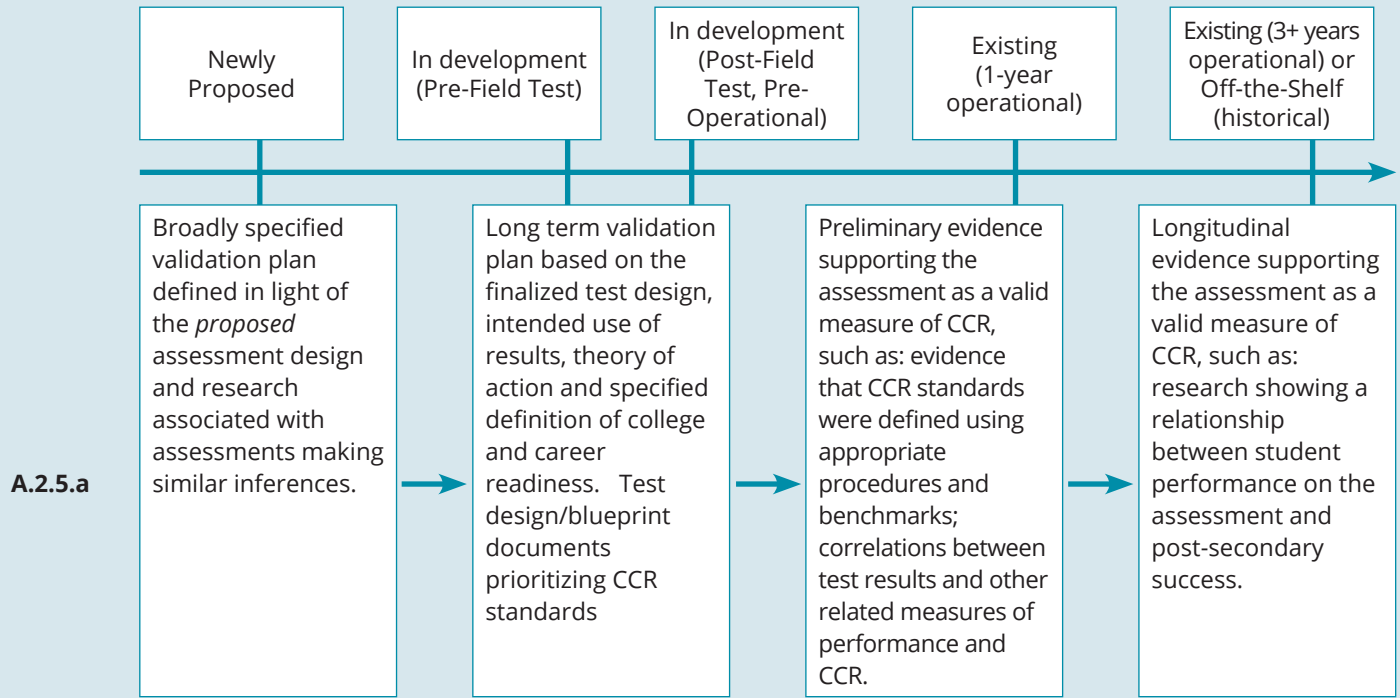
To illustrate this point consider claim A.2.5a which reads, "Evidence is provided to support the use of assessment results for making valid inferences about student performance and readiness for college and career (or on-track to CCR)." As shown in Figure 6, the type, amount, and sources of evidence one would expect to support this claim differs as you move along the assessment development continuum.

While the CEF includes comments intended to encourage consideration of phase of development or implementation when evaluating evidence (where appropriate), sufficiency statements generally reflect expectations associated with an existing assessment program. Therefore, the primary purpose of this discussion is to encourage evaluators to consider this factor when making and recording claim and criterion-level ratings.



TRAINING NOTES:

FIGURE 6. ASSESSMENT DEVELOPMENT CONTINUUM



While the CEF includes comments intended to encourage consideration of phase of development or implementation when evaluating evidence (where appropriate), sufficiency statements generally reflect expectations associated with an existing assessment program. Therefore, the primary purpose of this discussion is to encourage evaluators to consider this factor when making and recording claim and criterion-level ratings.

Step 4: Repeat Step 3 as adapted for the remaining test characteristics criteria. Note: Criterion A.2 should always be reviewed **last** as judgments regarding the degree to which the evidence provided supports the criterion statement will rely on the entire body of submitted evidence, including evidence provided for each of the other criteria.

During the review process evaluators should take comprehensive notes on the rating form so they can share their thinking with the Evaluation Team during Phase 3. Specifically if there were:

- particular pieces of evidence that weighed more heavily than others when evaluating a given claim;
- additional pieces of evidence (or claims) from the evidence list considered during review (outside those associated with the primary and secondary claims);
- specific contextual factors that strongly influenced the evaluation/rating.



TRAINING NOTES:

1. During training it should be stressed that the evaluators work independently. They should not have conversations with other members of the Evaluation Team about the evidence provided prior to the large group discussion.
2. If each evaluator is reviewing all of the criteria we strongly recommend that evaluators wait to make a final, overall criterion-level rating until all evidence (across all criteria) has been reviewed.²³ Evaluators may make a preliminary criterion-level rating once the full set of primary claims has been reviewed if they understand that these ratings can and should be revisited once the full set of evidence has been reviewed.
3. Criterion A.5 refers explicitly to students with disabilities and English learners, it may be helpful for reviewers to consider capturing ratings and notes relative to each of these groups separately. See Appendix F for an alternate way to create the independent evaluator sheet, specific to Criterion A.5, to support this purpose.

²³ One reason for this is to ensure secondary claims are considered. However even for criteria not having secondary claims, evidence submitted for a one criterion may ultimately influence how one rates a previously reviewed criterion even if not aligned to that criterion on the Evidence Log.

Step 5: After evaluating the evidence associated with each criterion and commenting on all specified claims, briefly summarize your thoughts related to the sufficiency of the body of evidence provided for each criterion, including specific examples that may be shared with the larger group. These comments should be captured on the page of the Independent Evaluation Rating form labeled “Overall Criterion Rating for [X]”, as represented in Appendix E (page 53).

Due to identified relationships between criteria it is recommended that evaluators wait to conduct this holistic criterion-level evaluation until after all claims (across all criteria) have been reviewed and considered. A variety of factors will influence the manner in which claims are rated and weighted in the evaluation of a given criterion. Since these factors will vary from assessment to assessment, as will the relevance/importance of secondary claims, criterion-level ratings cannot be assigned in a formulaic matter (i.e., in consideration of pre-specified patterns or proportions of ratings across claims). For example, a pattern of claim-level ratings for A.3.1-A.3.3 may be judged as representing Limited evidence of quality in one case and Good evidence of quality in another due to factors related to the design of the assessment, test taking population, or the test’s phase of development/implementation.

To facilitate consistency in the rating process, an operational definition of each criterion rating category is provided in Table 12.

TABLE 12. CRITERION-LEVEL RATING DESCRIPTORS

RATING	
Weak	The body of evidence presented provides weak support for this criterion, as defined in the Criteria Evaluation Framework. The evidence reviewed consistently did not meet expectations given the assessment’s current phase of development/implementation, design, and specified purpose(s) and goals.
Limited	The body of evidence presented provides limited support for this criterion, as defined in the Criteria Evaluation Framework. While some of the evidence may have met expectations, key elements central to this criterion are missing or poor given the assessment’s current phase of development/implementation, design, and specified purpose(s) and goals.
Good	The body of evidence presented provides adequate support for this criterion as defined in the Criteria Evaluation Framework. The preponderance of evidence reviewed met expectations given the assessment’s current phase of development or implementation, design, and specified purpose(s) and goals.
Excellent	The body of evidence presented provides strong support for this criterion, as defined in the Criteria Evaluation Framework. The evidence reviewed consistently met or exceeded expectations given the assessment’s current phase of development or implementation, design, and specified purpose(s) and goals. The evidence could serve as a strong exemplar or model for similar assessments.

Finally, given the descriptors provided in Tables 11 and 12, the following rules should hold when assigning criterion-level ratings.

Rules for Assigning Criterion Level Ratings

- If all of the primary claims associated with a given criterion are rated as having evidence that “Does Not Meet” expectations, the evidence associated with that criterion should receive an overall rating of **Weak**.
- If any of the primary claims associated with a criterion receive a rating of “Does Not Meet”, the evidence associated with that Criterion cannot be rated as **Excellent**.
- If all primary claims are rated as “Meets,” the evidence associated with that Criterion should receive an overall rating of **Good** or **Excellent**.



TRAINING NOTES:

- *Provide guidance that helps evaluators distinguish among the performance levels and allow ample time for group discussion.*
- *Encourage reviewers to think about evidence holistically rather than piece-by-piece when making criterion-level determinations.*
- *For those criteria that have multiple categories of claims, as indicated within the criteria explanation associated with each criterion in the CEF (i.e., A.1, A.2, A.4, A.7, D.2 and E.1), highlight the fact that it may be easier to consider the body of evidence associated with each category of claims before making an overall criterion-level rating.*

Step 6: Submit your Independent Evaluator Rating forms and any additional review comments to the Implementation Team.

Calibrate evaluators on the claim-level rating process and scale

As the last stage of the training process, it is recommended that the Implementation Team have evaluators review, discuss and rate one or two claims in order to calibrate their thinking about the rating process before independent evaluation. In order for this step to add value, the evidence associated with the claim(s) to be reviewed as a group should be provided to the Evaluation Team in advance of the meeting with explicit directions for review. The Implementation Team should facilitate this discussion by asking each evaluator to share his or her overall thoughts regarding the pool of evidence provided for a given claim and the rationale for their preliminary claim-level rating. Evaluators should be encouraged to discuss any contextual factors that influenced their decision, (e.g., age; CAT/non-CAT; etc.).

This activity is intended to establish a clear, common understanding of the expectations underlying the claim-level rubric so that it can be consistently applied across evaluators. To support this process, and better operationalize the levels, the Evaluation Team may be asked to provide examples that demonstrate how the evidence associated with a given claim might differ from one level to the next. Essentially answering the question “What would move this evidence from a category of Partially Meets to Meets, or Does Not Meet, to Partially Meets?”

Independent Evaluator Review

The second part of this phase of the evaluation methodology is when, after training, the Evaluators independently review and evaluate the evidence submitted on behalf of an assessment. This process can follow the six steps outlined in the section above, or follow some other protocol as determined and communicated by the Implementation team.

Phase 3: Group Discussion and Summary of Evaluation Findings

Shortly after the completion of Phase 2, the Evaluation Team should be convened for group discussion. The goal of this meeting is two-fold: 1) to establish a consensus rating on the strength of the evidence presented for each criterion and 2) articulate the components of an evidence-based argument for each rating that references the evidence reviewed in relation to the specified claims and relevant contextual factors.

To make Phase 3 as efficient as possible, the Implementation Team should review the full set of Phase 2 comments and ratings to identify key areas of disagreement and summarize comments and recommendations associated with perceived areas of consensus. This pre-work will permit facilitators from the Implementation Team to focus group discussion on areas of concern and orient the Evaluation Team to the group’s position prior to the meeting. In addition, proposed text summarizing the rationale and argument for perceived areas of consensus can be provided to the Evaluation Team for discussion and modification at the meeting (or beforehand as pre-meeting reading), rather than being generated from scratch.

To allow ample time for these activities, it should be assumed that this meeting will take at least two days; however, the Implementation Team may decide to shorten or extend this meeting as appropriate given the scope of the materials reviewed and the degree of agreement observed in Phase 2.

The Implementation Team is responsible for facilitating the discussion as well as taking detailed notes. Clearly, much more will be said than can make it into the final reports. It is the responsibility of the Implementation Team to capture and summarize consensus opinions and comments for inclusion in the final reports. To this end, the Implementation Team may decide to record the large group discussion so that it can be referenced as needed during report generation. Throughout the discussion



TRAINING NOTES:

- *This part of the training is extremely important so the Implementation Team should clarify that Evaluators are expected to review the evidence associated with the identified “training claim” prior to the training session.*
- *When selecting the training claim, try to pick one having enough evidence to support a comprehensive discussion, but not too much that it will take a whole day for an evaluator to review. The evidence should be one which allows for evaluators to clearly point to important features or missing elements that would influence their claim-level ratings.*
- *Reiterate that claim and criterion-level ratings should be based upon the expectations defined in the CEF in conjunction with relevant contextual factors such as the assessment’s phase of development. If discussion indicates that decisions are being based in large part on previous experience (i.e., what evaluators typically see in reference to a given claim) be sure to call this out and make it a point of discussion.*

process, comments regarding the extent to which the reviewed assessments meet the CCSSO Criteria and specific areas of strength and weakness should be documented. When consensus views and ratings cannot be established (at the claim or criterion level), minority views should also be noted for inclusion in the final report.²⁴

While a proposed structure for the large group discussion is outlined in Table 13, the Evaluation Team may come up with its own strategy to achieve the desired end-goal. The Implementation Team is charged with making sure that the evaluators stay on-track and review each criterion, but the process used to get there can be defined, in large part, by the evaluators.

TABLE 13. PROPOSED ELEMENTS OF THE EVALUATION TEAM MEETING

<ol style="list-style-type: none"> 1. Group Introductions (if necessary). 2. The Implementation Team outlines the goal of the meeting and describes, in detail, what needs to be accomplished. 3. Each evaluator is provided with 5 minutes to summarize their thoughts related to the evaluation process and the pool of evidence provided. 4. A facilitator from the Implementation Team summarizes the Phase 2 findings regarding general areas of agreement/disagreement and proposes a strategy for the discussion in light of these results. (For example, if the ratings and comments associated with A.3 suggest that the evaluators considered the evidence associated with this criterion differently and/or focused on different pieces of evidence during their review, this may be a good place to begin.) 5. The evaluators accept or modify this plan as they believe appropriate, and begin their discussions of each criteria. 6. To establish an indication of the degree to which there is variability in thinking, evaluators should initiate the discussion of a particular criterion by sharing their preliminary ratings of the overall quality of evidence presented (e.g., Weak, Limited, Good, Excellent). Subsequently, the Implementation Team should facilitate a discussion focused on identifying those claims and factors that had the greatest impact on the evaluators' proposed ratings. 7. After discussing each claim associated with a criterion the group should come to consensus on: 1) the overall rating to be assigned to that criterion, and 2) the key factors influencing that decision which should be highlighted in the final report. To the extent possible, contextual factors and relationships between criteria that were critical to the evaluation should be noted. 8. If a draft argument or explanation has been developed by the Implementation Team for a given criterion (i.e., in light of their initial review of the Phase 2 results), this should be presented to the Evaluation Team for review and modification.



TRAINING NOTES:

- *As evaluators respond and discuss as a group, the Implementation Team should be informing the discussion by providing the preliminary ratings and comments summarized as a result of independent evaluator review.*
- *Some possible conversation starters for the large group discussion of each criterion include:*
 1. *Were there any claims, or categories of claims, for which the evidence provided clearly “Did Not Meet” expectations? If so, why? (Note: If yes, this criterion cannot receive a rating of Excellent)*
 2. *Were there any claims, or categories of claims, for which the evidence provided far exceeded expectations? If so, why?*
 3. *What are your biggest concerns about the body of evidence provided for this criterion? What key elements were missing? What type of evidence would serve to fill those gaps?*
 4. *What evidence lent the strongest support to this criterion?*
 5. *How did you prioritize/weight the different claims or categories of claims in making your overall recommendation for this criterion? Did certain claims hold more/less weight, or were all claims equally important in this context?*
 6. *Did the secondary claims influence your rating, if so in which way? What, if any, additional claims should be considered when evaluating this criterion?*
 7. *Given the observed strengths and weakness and the assessment’s current phase of development or implementation, how should we rate the overall body of evidence provided to support this criterion?*

²⁴ To the extent possible the Implementation Team should strive to establish consensus with respect to criterion-level ratings.

Phase 4: Report Generation and Approval

Two evaluation reports will serve as the final products of the test characteristics evaluation: 1) a high-level **executive report**, and 2) a detailed **claim-level evaluation report**. The first report documents the consensus rating for each criterion along with a brief summary of the rationale for that rating. The executive report should also clearly indicate the assessment's phase of development/implementation at the time of evaluation. This report is should be created by the Implementation Team after the completion of Phase 3. A sample template for this executive report can be found in Appendix G. The format of this template is analogous to one of the final reporting products produced by the test content evaluation and can be put side-by-side to create a comprehensive evaluation report for the assessment that spans all of the CCSSO Criteria.

The second report provides information about each criterion as well as comments summarizing the quality of evidence reviewed at the claim level. This report is intended for a more technical audience and should be designed to supplement and reinforce the evaluation results summarized in the executive report. The Implementation Team should begin drafting this report after the completion of independent evaluator review (i.e., Phase 2) and prior to the large group discussion (Phase 3). This will allow the Evaluation Team to review and modify the preliminary report during Phase 3. Additionally, the report can serve as a starting point for conversation. After Phase 3, the Implementation Team should revise the initial draft of the evaluation report, as needed, based on feedback and comments provided by the Evaluation Team.

The format of the evaluation report can be defined by the Implementation Team, but it should include at least the following information: purpose of the evaluation, requesting organization, members of the evaluation team, evidence log, prioritized evidence list, the executive summary report which provides the consensus rating for each criterion and its evidence-based rationale (see Appendix E, page 53), and a comprehensive report, which provides addition detail regarding the quality of evidence presented for each criterion (i.e., summaries of comments and concerns associated with each claim. The degree of consensus among evaluators in making claim-level ratings is not necessary to include in the comprehensive report as long as evaluators were able to come to a shared understanding about the body of evidence and agree upon criterion-level ratings. Upon completion of the intermediate draft, the evaluation report should be provided to the Evaluation Team for review and comment before finalization and release to the requesters. Once the report has been completed, the Providers should be given the opportunity to review and, if desired, provide a brief explanation or set of dissenting comments for consideration in an appendix to the report.

Considerations Related to the Implementation of the Methodology for Assessments Developed for Use in Multiple States

Those charged with implementing the evaluation of a test developed for use in multiple states will need to carefully consider what evidence can and should be collected to support evaluation and, consequently, the types of statements/ratings that can be made to support inferences about the quality of the test. Where appropriate and feasible, the CEF identifies unique considerations related to the evaluation of summative assessments aligned to college and career ready standards that are intended to serve multiple states. Unfortunately, all such considerations cannot be anticipated, and those which tend to be most influential will likely vary from test-to-test. A series of questions intended to support the Implementation Team in thinking this through for each of the CCSSO Criteria is provided below.

1. Can all of the claims associated with this criterion be evaluated using a common, unitary set of evidence? That is, would the evidence provided to support each claim be the same for all states administering the assessment?
2. For those claims that cannot be addressed using common evidence, why? What is the nature of the variability that might be observed across states?
3. For those claims that cannot be fully addressed using common evidence, are there particular *components* that might be based on the same evidence for all states, supporting a partial evaluation? (For example, does the consortium/assessment vendor provide a common set of guidelines to support scoring and reporting of the assessments even if different vendors are selected to apply these guidelines across states?)
4. Given the number of claims for which common evidence will be available, will there be enough information by which to make a valid, overall rating regarding the extent to which this CCSSO criterion (as defined) has been met? (That is, to make a rating which you believe to be appropriate for and generalizable to all states utilizing this assessment?)

It is assumed that the Implementation Team and Requester can work through these questions independently; however, if necessary, support may be requested from a Provider or member of the Evaluation Team. Once collected, the Implementation Team can use the responses to these questions to inform decisions about how to conduct the evaluation and report results in a

manner that provides for useful and accurate information. While there is no one right or best way, as contextual factors related to the intended purpose and use of evaluation results will always come into play, a few options for consideration are provided below:

- A. If all of the criteria can be evaluated using common evidence (i.e., the answer to question #1 is “yes”), conduct the evaluation and report the final rating associated with each claim and criterion without caveats.
- B. If there are some criteria that can only be partially evaluated with common evidence, but the decision is not to gather state-specific evidence due to time constraints, resources or appropriateness, options include:
 - a. Divide the evaluation into two parts – Part 1 would include those criteria which can be fully evaluated with common evidence and Part 2 would include those criteria which require state-specific evidence, or a combination of state-specific and common evidence to support the provision of a criteria-based rating. Part 1 could be conducted and reported for criteria based on common evidence, with the understanding that Part 2 would require the collection and evaluation of evidence specific to each state in order to make a final rating on these criteria. Within the context of the Part 2 evaluation, specific claims evaluated using common evidence across states should be identified.
 - b. Have evaluators make ratings for all claims that can be supported with common evidence, even if associated with a criterion that requires state-specific evidence to be fully evaluated. For those criteria that require state-specific evidence in addition to common evidence the final report would include ratings and rationales *only at the claim level* and clearly indicate the types of statements that can/cannot be supported given the common evidence reviewed.
- C. If there are criteria that can only be partially evaluated with common evidence, but it is appropriate and possible to gather a sample of state-specific evidence that informs evaluation, options for consideration include the following:
 - a. If a claim requires state evidence related to implementation, but is based upon a set of consortium (or vendor) developed guidelines, the Implementation Team may decide to compile a sample of evidence from multiple states so evaluators can determine the *degree to which guidance provided by the consortia/vendor is supporting adherence to claims at the statelevel*.
 - b. If there is a particular criterion for which the majority of evidence provided is consortium/vendor-based or common across states, the Evaluation Team may provide an overall rating for this criterion with caveats clearly noted in terms of what (if any) specific claims may not be fully supported and/or any conditions which must hold for the rating to fully generalize.
- D. If there are some criteria that can only be partially evaluated with common evidence, but the interest is in evaluating the administration of the assessment ***in just one state***, collect state-specific evidence to support a complete evaluation. The report should clearly indicate that the evaluation is specific to the administration of the assessment in that state only and the results, therefore, do not generalize.

With respect to bullet b under option C, consider criterion A.4, which addresses the development and review of test items, assessments and score scales. If all states are using the same operational form or test bank, then evidence associated with the development and review of test items and forms, for example, would be common to all states. If however, a state decides to augment an assessment developed for use in multiple states with its own items this evidence is no longer fully generalizable. In this case, the final report should include caveats that inform the manner in which this claim can be interpreted. For example, “This applies only to those items developed using the vendor’s procedures. If a test developed for use in multiple states is modified with state-specific items, general claims regarding the quality of those items cannot be made without additional evidence related to the state-specific item development and review procedures.”

Clearly there are implications associated with each of the options provided above which will need to be weighed and considered when defining the best way to proceed. There are many reasonable ways to conduct high quality evaluations of assessments developed for multiple states; however, one must be careful to sort through options and choose an approach that best meets needs of those requesting the evaluation. The options provided above are just one way to think about how one might conduct and organize this type of review.

Considerations Related to the Implementation of the Methodology for Multiple Assessments Simultaneously

The procedures outlined in the previous sections apply mainly to those situations where one test has been selected or identified for evaluation by a requesting organization. In some cases, however, it may be necessary or desirable to evaluate multiple assessment programs within a similar timeframe. For example, a Requester may be interested in supplying states with descriptive information about the quality of evidence available for several common (i.e., high use) assessment programs so that states administering those assessments understand where evidence of quality is lacking (e.g., to support peer review, or validate a particular use of assessment results). In other cases, a requester may be interested in comparing evaluation results across multiple assessments with the goal of selecting the assessment which provides for a profile of evidence that is best aligned to their needs.

In the latter scenario, the goal of evaluation is not just to provide descriptive information about the quality of evidence available, but to compare results across assessments in service to a specific goal or use. In this case, consistency in the evaluation and reporting process should be built into the design of the methodology to the greatest extent possible. One of the easiest ways to do this is to use the same set of evaluators for all of the assessments to be reviewed. Using the same Evaluation Team promotes consistency in the interpretation of the CEF and the rating of claims and criteria across assessment programs. While consistency can be established, to some extent, through the use of a detailed training process, having the same panel comment on different assessments using a consistent language, format and argument structure will help to facilitate the comparison of results.

The use of common evaluators across assessments can, however, pose some logistical challenges. This is especially true if the amount of evidence to be reviewed is large and the timeline for implementation short. While we strongly believe, for reasons previously discussed, it is optimal for each evaluator to review all of the evidence (across criteria) associated with a given program, we understand that this may not always be feasible. In this case, a matrix-based evaluation approach might be considered. Under a matrix approach, each evaluator is assigned to review and comment on a common subset of the CCSSO Criteria across all the assessment programs to be reviewed. For example, in a simple case where there are two assessment programs to be considered, Evaluator 1 may be asked to review the evidence associated with criteria A.1, A.2 and A.4 in assessment programs 1 and 2 so that there is much consistency in the evaluation and rating process of these criteria as possible. When there are more than two programs under review the evaluation design will clearly be more complex, requiring either more evaluators (e.g., each reviewing 2-3 “common” criteria for each assessment) and/or a longer evaluation window to accommodate the increased number of programs to be reviewed.

If a matrix-based design is considered, the Implementation Team must be extremely thoughtful with respect to 1) how criteria are chunked together for review, and 2) how sets of criteria are assigned to members of the Evaluation Team. For many criteria, both primary and secondary claims have been identified to highlight some of the relationships and dependencies between criteria. Since evaluators are encouraged to consider the evidence associated with both primary and secondary claims when making an overall criterion rating, to the extent possible, criteria having related sets of claims should be chunked together for assignment. For example criterion A.1, has secondary claims from A.4; A.3 includes claims from D.1, A.4 has secondary claims from E.1 and A.5; and A.7 includes claims from E.1.²⁵

In addition to the specified secondary claim, the Implementation Team should be cognizant of what information about the assessment will be necessary for evaluators to know in order to make judgments about a limited set of claims. For example, in order to make judgments about the content and format of score reports (claim D.1.1), information regarding the scaling will be prerequisite knowledge. Furthermore, since the methodology recommends that overall ratings for A.2 (validity) take into account the evidence associated with each of the other criteria (as secondary), we recommend that all evaluators be assigned A.2 for review so a large group discussion of this criterion can occur. An additional consideration to take into account is the overlap in the evidence associated with different criteria. For many criteria, a subset of the artifacts identified to support evaluation will be the same (e.g., an annual technical report, or the program’s theory of action). To support efficiency and coherence in the evaluation process, to the extent possible criteria sharing common evidence should be chunked together for review.

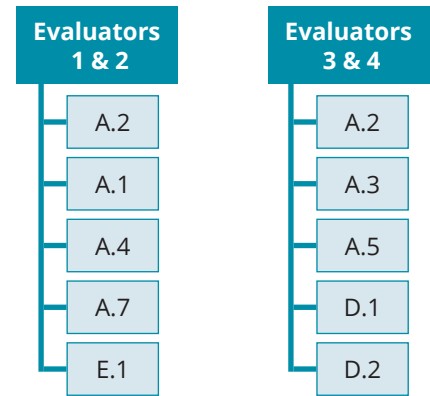
Finally, since the Implementation Team is encouraged to select a team of evaluators that represents different areas of focal expertise, such expertise must be taken into account when assigning evaluators to criteria. For example, an evaluator with expertise in the area of developing and analyzing assessments for students with disabilities should be assigned A.5 and any related criteria (e.g., A.2, E.1). Figure 7 shows one way to group the criteria that may make sense; however, as previously stated, these groupings should be thoughtfully considered based on the assessment programs reviewed, the evidence provided, and the expertise of the Evaluation Team.

²⁵ It is important to note that while primary and secondary claims have been identified, an Evaluation Team may identify other cross-criterion dependencies they believe to be important for consideration. One short-coming of the matrix-based approach is that some of these dependencies may be missed if each evaluator is not reviewing all of the evidence associated with a given assessment.

While most of the procedures outlined for independent and group-based review will still be appropriate under a matrix-based approach, a few modifications will likely be necessary.

- If there is a large amount of evidence to be reviewed across the programs, additional evaluators may be necessary. The evidence associated with each criterion (except A.2) should be reviewed by **at least** two evaluators so that discussions about ratings and their rationale can occur. A.2 should be reviewed by each member of the Evaluation Team.
- Because each evaluator will not be reviewing all of the evidence associated with a given program, evidence that is crucial to the overall interpretation of the assessment program that cannot be sufficiently represented in training should be assigned in advance to each member of the entire Evaluation Team, even if not specifically aligned to his/her assigned criteria (e.g., theory of action, test design documentation, purpose/use statements, etc..)
- During Phase 3, large group discussion and evaluation, those evaluators charged with the review of common criteria should meet to discuss and outline a rationale for their claim and criteria-level ratings. Subsequently, the entire Evaluation Team should meet to discuss the evidence as a whole and determine an appropriate rating for A.2.
- Although each evaluator will be reviewing multiple programs, a separate evaluation process should be conducted for each assessment. That is, the evidence associated with A.1 for assessment programs 1 and 2 should not be evaluated concurrently (i.e., side-by-side). Doing so increases the likelihood that ratings will be made on a relative basis rather than in considering key contextual factors unique to each test (e.g., phase of development or implementation, theory of action, assessment design, etc.). For each assessment there should be a separate training that provides information about the context of the assessment and evidence submitted for evaluation, an independent evaluator review (for assigned criteria), a large group discussion and a final report. That is, each program should be evaluated in turn rather than simultaneously.²⁶

FIGURE 7: POSSIBLE DIVISION OF CRITERIA



It is important to note that even when procedures are put in place to facilitate comparisons among tests, evaluation results must be considered in context. Therefore, final reports should avoid presenting criterion-level ratings for different programs side-by-side in the absence of evaluator rationales or important context such as the stage of the development process. For example, two assessments in different stages of development may receive the same overall rating for a criterion A.1 (which requires evidence related to how scores are mapped to determinations of college and career readiness) despite providing completely different types of evidence. In one case evidence may include proposed procedures or materials/outcomes associated with assessments similar to that under evaluation (i.e., if the assessment is newly proposed); while in the other it may include the materials and outcomes resulting from implementation (i.e., if the assessment already exists). If ratings are reviewed in the absence of evaluator rationales this important distinction may be overlooked.

If the primary goal of evaluation is to compare assessments for the purpose of selecting one for use, the requester may consider adding an additional phase to the evaluation process in which the Evaluation Team discusses the pros/cons of each program given the requesters stated goals and uses for assessment and provides an overall recommendation. This final phase should be summarized in a separate section with distinct and separate findings and recommendations.

CONTACT

Though this methodology has undergone a number of internal and external reviews including a test case, we acknowledge that there may be errors or oversights. As you review this documentation and hopefully use it to support an operational evaluation of an assessment program we encourage you to reach out to us with your feedback and commentary. For this purpose we have included the contact information of the primary developers below. Thank you.

Erika Hall, Ph.D.
ehall@nceia.org

Susan Lyons, Ph.D.
slyons@nceia.org

²⁶ While each program should be evaluated in turn, counter-balancing the order in which assessment programs are reviewed across programs may be desirable to decrease the likelihood of an order effect; however this may require deferring the large group discussion until after training and independent evaluator review of each assessment program has occurred.

APPENDICES

- A. Crosswalk Between CCSSO Evidence Statements and Criteria Evaluation Framework
- B. Complete Structure of Claims for Evaluation of Test Characteristics
- C. Examples of Evidence that May be Provided for Evaluation
- D. Templates to Support the Collection and Organization of Evidence
- E. Independent Evaluation Rating Sheet
- F. Independent Evaluator Rating Sheet Sample Specific to Criterion A.5
- G. Sample Test Characteristics Summary Report Template

APPENDIX A: CROSSWALK BETWEEN CCSSO EVIDENCE STATEMENTS AND CRITERIA EVALUATION FRAMEWORK

A. MEET OVERALL ASSESSMENT GOALS AND ENSURE TECHNICAL QUALITY²⁷

CRITERIA	EVIDENCE	CROSSWALK TO CEF
A.1 Indicating progress toward college and career readiness: Scores and performance levels on assessments are mapped to determinations of college and career readiness at the high school level, and for other grades, to being on track to college and career readiness by the time of high school graduation	All eleven claims associated with Criterion A.1 are logical extensions of the language expressed in the original CCSSO document and are designed to specify a comprehensive set of evidence that supports evaluation. The specific claims listed in column 3 below indicate where each of the evidence statements presented in the CCSSO documentation can be located within the Criteria Evaluation Framework.	
	<ul style="list-style-type: none"> • A description is provided of the process for developing performance level descriptors and setting performance standards (i.e., “cut scores”), including: <ul style="list-style-type: none"> - Appropriate involvement of higher education and career/technical experts in determining the score at which there is a high probability that a student is college and career ready; 	A.1.2, A.1.6
	<ul style="list-style-type: none"> - External evidence used to inform the setting of performance standards and a rationale for why certain forms of evidence are included and others are not (e.g., student performance on current state assessments, NAEP, TIMSS, PISA, ASVAB, ACT, SAT, results from Smarter Balanced and PARCC, relevant data on post-secondary performance and remediation and workforce readiness); 	A.1.8
	<ul style="list-style-type: none"> - Evidence and a rationale that the method(s) for including external benchmarks are valid for the intended purposes; and 	A.1.8
	<ul style="list-style-type: none"> - Standard setting studies, the resulting performance level descriptors and performance standards, and the specific data on which they are based (when available). 	A.1.2, A.1.10, A.1.11 *Final PLDs and cut scores are not directly reviewed
<ul style="list-style-type: none"> • A description is provided of the intended studies that will be conducted to evaluate the validity of performance standards over time. 	A.1.10	

²⁷ The term “technical quality” here refers to the qualities necessary to ensure that scoring and generalization inferences based on test scores are valid both within and across years. This document prioritizes certain aspects of technical quality, but as noted in the introduction, readers should also refer to other sources, primarily *The Standards for Educational and Psychological Testing*.

CRITERIA	EVIDENCE	CROSSWALK TO CEF
<p>A.2 Ensuring that assessments are valid for required and intended purposes: Assessments produce data, including student achievement data and student growth data required under Title I of ESEA and ESEA Flexibility, that can be used to validly inform the following:</p> <ul style="list-style-type: none"> • School effectiveness and improvement; • Individual principal and teacher effectiveness, for purposes of evaluation and identification of professional development and support needs; • Individual student gains and performance; and • Other purposes defined by the state. 	<p>All seven claims associated with Criterion A.2 are logical extensions of the language expressed in the original CCSSO document and are designed to specify a comprehensive set of evidence that supports evaluation. The specific claims listed in column 3 below indicate where each of the evidence statements presented in the CCSSO documentation can be located within the Criteria Evaluation Framework.</p>	
	<ul style="list-style-type: none"> • A well-articulated validity evaluation based on an interpretive argument (e.g. Kane 2006) is provided that includes, at a minimum: <ul style="list-style-type: none"> - Evidence of the validity of using results from the assessments for the three primary purposes, as well as any additional purposes required by the state (specify sources of data). 	A.2.5
	<ul style="list-style-type: none"> - Evidence that scoring and reporting structures are consistent with structures of the state's standards (specify sources of data). 	A.4.6, D.1.1
	<ul style="list-style-type: none"> - Evidence that total test and relevant sub-scores are related to external variables as expected (e.g., other measures of the construct). To the extent possible, include evidence that the items are "instructionally sensitive," that is, that item performance is more related to the quality of instruction than to out-of-school factors such as demographic variables. 	A.2.5.a, A.2.5.b, A.2.5.c
	<ul style="list-style-type: none"> - Evidence that the assessments lead to the intended outcomes (i.e., meet the intended purposes) and minimize unintended negative consequences. Consequential evidence should flow from a well-articulated theory of action about how the assessments are intended to work and be integrated with the larger accountability system. 	A.2.6
	<ul style="list-style-type: none"> - The set of content standards against which the assessments are designed is provided. If these standards are the state's standards, evidence is provided that the content of the assessments reflects the standards, including the cognitive demand of the standards. If they are not the state's standards, evidence is provided of the extent of alignment with the state's standards. 	A.2.5.a.
	<ul style="list-style-type: none"> - Evidence is provided to ensure the content validity of test forms and the usefulness of score reports (e.g., test blueprints demonstrate the learning progressions reflected in the standards, and experts in the content and progression toward readiness are significantly involved in the development process). 	A.2.5.a., A.4.5, D.1.2

CRITERIA	EVIDENCE	CROSSWALK TO CEF
<p>A.3 Ensuring that assessments are reliable: Assessments minimize error that may distort interpretations of results, estimate the magnitude of error, and inform users of its magnitude.</p>	<p>All three claims associated with Criterion A.3 are logical extensions of the language expressed in the original CCSSO document and are designed to specify a comprehensive set of evidence that supports evaluation. The specific claims listed in column 3 below indicate where each of the evidence statements presented in the CCSSO documentation can be located within the Criteria Evaluation Framework.</p>	
	<ul style="list-style-type: none"> Evidence is provided of the reliability of assessment scores, based on the state’s student population and reported subpopulations (specify sources of data). 	A.3.3
	<ul style="list-style-type: none"> Evidence is provided that the scores are reliable for the intended purposes for essentially all students, as indicated by the standard error of measurement across the score continuum (i.e., conditional standard error). 	A.3.3
	<ul style="list-style-type: none"> Evidence is provided of the precision of the assessments at cut scores, and consistency of student level classification (specify sources of data). 	A.3.3
	<ul style="list-style-type: none"> Evidence is provided of generalizability for all relevant sources, such as variability of groups, internal consistency of item responses, variability among schools, consistency from form to form of the test, and inter-rater consistency in scoring (specify sources of data). 	A.3.3
<p>A.4 Ensuring that assessments are designed and implemented to yield valid and consistent test score interpretations within and across years:</p> <ul style="list-style-type: none"> Assessment forms yield consistent score meanings over time, forms within year, student groups, and delivery mechanisms (e.g., paper, computer, including multiple computer platforms). Score scales used facilitate accurate and meaningful inferences about test performance. 	<p>All eleven claims associated with Criterion A.4 are logical extensions of the language expressed in the original CCSSO document and are designed to specify a comprehensive set of evidence that supports evaluation. The specific claims listed in column 3 below indicate where each of the evidence statements presented in the CCSSO documentation can be located within the Criteria Evaluation Framework.</p>	
	<ul style="list-style-type: none"> A description is provided of the process used to ensure comparability of assessments and assessment results across groups and time. 	A.4.4
	<ul style="list-style-type: none"> Evidence is provided of valid and reliable linking procedures to ensure that the scores derived from the assessments are comparable within year across various test “forms” and across time. 	A.4.9
	<ul style="list-style-type: none"> Evidence is provided that the linking design and results are valid for test scores across the achievement continuum. 	A.4.10
	<ul style="list-style-type: none"> Evidence is provided that the procedures used to transform raw scores to scale scores is coherent with the test design and the intended claims, including the types of IRT calibration and scaling methods (if used) and other methods for facilitating meaningful score interpretations over tests and time. 	A.4.7
	<ul style="list-style-type: none"> Evidence is provided that the assessments are designed and scaled to ensure that the primary interpretations of the assessment can be fulfilled. For example, if the assessments are used as data sources for growth or value-added models for accountability purposes, evidence should be provided that the scaling and design features would support such uses, such as ensuring appropriate amounts of measurement information throughout the scale, as appropriate. 	A.4.6
	<ul style="list-style-type: none"> Evidence is provided, where a vertical or other score scale is used, that the scaling design and procedures lead to valid and reliable score interpretations over the full length of the scale proposed; and evidence is provided that the scale is able to maintain these properties over time (or a description of the proposed procedures is provided). 	

CRITERIA	EVIDENCE	CROSSWALK TO CEF
A.5 Providing accessibility to all students, including English learners and students with disabilities:	All four claims associated with Criterion A.5 are logical extensions of the language expressed in the original CCSSO document and are designed to specify a comprehensive set of evidence that supports evaluation. The specific claims listed in column 3 below indicate where each of the evidence statements presented in the CCSSO documentation can be located within the Criteria Evaluation Framework.	
<ul style="list-style-type: none"> • 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. 	<ul style="list-style-type: none"> • A description is provided of the item development process used to reduce construct irrelevance (e.g., eliminating unnecessary clutter in graphics, reducing construct-irrelevant reading load as much as possible), including <ul style="list-style-type: none"> - The <i>test item</i> development process to remove potential challenges due to factors such as disability, ethnicity, culture, geographic location, socioeconomic condition, or gender; and - <i>Test form</i> development specifications that ensure that assessments are clear and comprehensible for all students. • Evidence is provided, including exemplar tests (paper and pencil forms or screen shots) illustrating principles of universal design. 	A.5.1
<ul style="list-style-type: none"> • 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. 	<ul style="list-style-type: none"> • A description is provided of the accessibility features that will be available, consistent with state policy (e.g., magnification, audio representation of graphic elements, linguistic simplification, text-to-speech, speech-to-text, Braille). • A description is provided of access to translations and definitions, consistent with state policy. • A description is provided of the construct validity of the available accessibility features with a plan that ensures that the scores of students who have accommodations or modifications that do not maintain the construct being assessed are not combined with those of the bulk of students when computing or reporting scores. 	A.5.2
<ul style="list-style-type: none"> • Assessments produce valid and reliable scores for English learners. 	<ul style="list-style-type: none"> • Evidence is provided that test items and accessibility features permit English learners to demonstrate their knowledge and abilities and do not contain features that unnecessarily prevent them from accessing the content of the item. Evidence should address: presentation, response, setting, and timing and scheduling (specify sources of data). 	A.5.3, A.5.4
<ul style="list-style-type: none"> • Assessments produce valid and reliable scores for students with disabilities. 	<ul style="list-style-type: none"> • Evidence is provided that test items and accessibility features permit students with disabilities to demonstrate their knowledge and abilities and do not contain features that unnecessarily prevent them from accessing the content of the item. Evidence should address: presentation, response, setting, and timing and scheduling (specify sources of data). 	A.5.3, A.5.4

CRITERIA	EVIDENCE	CROSSWALK TO CEF
<p>A.7 Meeting all requirements for data privacy and ownership: All assessments must meet federal and state requirements for student privacy, and all data is owned exclusively by the state.</p>	<p>All six claims associated with Criterion A.7 are logical extensions of the language expressed in the original CCSSO document and are designed to specify a comprehensive set of evidence that supports evaluation. The specific claims listed in column 3 below indicate where each of the evidence statements presented in the CCSSO documentation can be located within the Criteria Evaluation Framework.</p>	
	<ul style="list-style-type: none"> • An assurance is provided of student privacy protection, reflecting compliance with all applicable federal and state laws and requirements. 	A.7.1
	<ul style="list-style-type: none"> • An assurance is provided of state ownership of all data, reflecting knowledge of state laws and requirements. 	A.7.4
	<ul style="list-style-type: none"> • An assurance is provided that state will receive all underlying data, in a timely and useable fashion, so it can do further analysis as desired, including, for example, achievement, verification, forensic, and security analyses. 	A.7.5
	<ul style="list-style-type: none"> • A description is provided for how data will be managed securely, including, for example, as data is transferred between vendors and the state. 	A.7.3, A.7.6

D. YIELD VALUABLE REPORTS ON STUDENT PROGRESS AND PERFORMANCE

CRITERIA	EVIDENCE	CROSSWALK TO CEF
<p>D.1 Focusing on student achievement and progress to readiness: Score reports illustrate a student's progress on the continuum toward college and career readiness, grade by grade, and course by course. Reports stress the most important content, skills, and processes, and how the assessment focuses on them, to show whether or not students are on track to readiness.</p>	<p>All three claims associated with Criterion D.1 are logical extensions of the language expressed in the original CCSSO document and are designed to specify a comprehensive set of evidence that supports evaluation. The specific claims listed in column 3 below indicate where each of the evidence statements presented in the CCSSO documentation can be located within the Criteria Evaluation Framework.</p>	
	<ul style="list-style-type: none"> • A list of reports is provided, and for each report, a sample that shows, at a minimum: <ul style="list-style-type: none"> - Scores and sub-scores that will be reported, with emphasis on the most important content, skills, and processes for each grade or course; 	D.1.1
	<ul style="list-style-type: none"> - Explanations of results that are instructionally valuable and easily understood by essentially all audiences; 	D.1.2, D.2.3
	<ul style="list-style-type: none"> - Results expressed in terms of performance standards (i.e., proficiency "cut scores"), not just scale scores or percentiles; and 	D.1.3
	<ul style="list-style-type: none"> - Progress on the continuum toward college and career readiness, which can be expressed by whether a student has sufficiently mastered the current grade or course content and is therefore prepared for the next level. 	D.1.3
	<ul style="list-style-type: none"> • The reporting structure can be supported by the assessment design, as demonstrated by evidence, including data confirming that test blueprints include a sufficient number of items for each reporting category, so that scores and subscores lead to the intended interpretations and minimize the possibility of misinterpretation. 	D.1.1
<p>D.2 Providing timely data that informs instruction: Reports are instructionally valuable, are easy to understand by all audiences, and are delivered in time to provide useful, actionable data to students, parents, and teachers.</p>	<p>All three claims associated with Criterion D.2 are logical extensions of the language expressed in the original CCSSO document and are designed to specify a comprehensive set of evidence that supports evaluation. The specific claims listed in column 3 below indicate where each of the evidence statements presented in the CCSSO documentation can be located within the Criteria Evaluation Framework.</p>	
	<ul style="list-style-type: none"> • A timeline and other evidence are provided to show when assessment results will be available for each report. 	D.2.2
	<ul style="list-style-type: none"> • A description is provided of the process and technology that will be used to issue reports in as timely a manner as possible. 	D.2.1
<ul style="list-style-type: none"> • Evidence, including results of user testing, is provided to demonstrate the utility of the reports for each intended audience. 	D.1.2	

E. ADHERE TO BEST PRACTICES IN TEST ADMINISTRATION

CRITERIA	EVIDENCE	CROSSWALK TO CEF
<p>E.1 Maintaining necessary standardization and ensuring test security: In order to ensure the validity, fairness, and integrity of state test results, the assessment systems maintain the security of the items and tests as well as the answer documents and related ancillary materials that result from test administrations.</p>	<p>All eight claims associated with Criterion E.1 are logical extensions of the language expressed in the original CCSSO document and are designed to specify a comprehensive set of evidence that supports evaluation. The specific claims listed in column 3 below indicate where each of the evidence statements presented in the CCSSO documentation can be located within the Criteria Evaluation Framework.</p>	
	<ul style="list-style-type: none"> • A comprehensive security plan is provided with auditable policies and procedures for test development, administration, score reporting, data management, and detection of irregularities consistent with NCES and CCSSO recommendations for, at a minimum: <ul style="list-style-type: none"> - Training for all personnel – both test developers and administrators; 	<p>E.1.3, E.1.4</p>
	<ol style="list-style-type: none"> 1. Secure management of assessments and assessment data, so that no individual gains access to unauthorized information; 	<p>E.1.3</p>
	<ol style="list-style-type: none"> 2. Test administration and environment; and 	<p>E.1.1, E.1.2</p>
	<ol style="list-style-type: none"> 3. Methods used to detect testing irregularities before, during and after testing, and steps to address them. 	<p>E.1.7, E.1.8</p>
	<ul style="list-style-type: none"> • A description is provided of how security safeguards have been tested and validated for computer-based tests and for paper-and-pencil tests, as relevant. 	<p>E.1.5</p>

APPENDIX B: COMPLETE STRUCTURE OF CLAIMS FOR EVALUATION OF TEST CHARACTERISTICS

A.1 Indicating progress toward college and career readiness: Scores and performance levels on assessments are mapped to determinations of college and career readiness at the high school level and for other grades being on track to college and career readiness by the time of high school graduation.

Primary claims related to the definition of CCR:

- A.1.1. College- and career readiness has been clearly defined for operational use.

Primary claims related to performance level descriptors:

- A.1.2. The process for developing performance level descriptors (PLDs) provides for PLDs that accurately represent the expectations defined by the CCR content standards within and across grades.
- A.1.3. Knowledgeable experts were involved in the process of developing and reviewing the PLDs.
- A.1.4. The process used for developing performance level descriptors (PLDs) supports their intended use(s).
- A.1.5. The process for developing performance level descriptors (PLDs) includes an evaluation of alignment of the PLDs to the content of the test questions that differentiate performance at each level, and, as needed, re-writing based on new evidence concerning skills needed for success in college and careers.

Primary claims related to standard setting:

- A.1.6. A description and coherent rationale are provided for how the proposed and/or implemented standard setting methodology²⁸ yields valid determinations of progress toward, or attainment of, college and career readiness.
- A.1.7. A coherent rationale accompanies methodological decisions regarding the level of involvement of grade-level educators, higher education, industry, and career technical experts (CTEs) in the standard setting process.
- A.1.8. Appropriate external CCR benchmarks and research studies are/were used in the standard setting process.
- A.1.9. Procedures and rationales for any adjustments made to proposed cut scores after the standard setting meeting are based on a defensible rationale and method.
- A.1.10. Studies planned or conducted to evaluate the validity of CCR performance standards over time are appropriate given the inferences they are intended to support.
- A.1.11. The standard setting procedures were followed as specified, and the final cut scores and the results of validity studies have been reviewed by technical experts.

Secondary claims from A.4 related to scaling: A.4.6 – A.4.10

A.2 Ensuring that assessment results are valid for required and intended purposes: Assessments produce student achievement and student growth data, as required under Title 1 of the Elementary and Secondary Education Act (ESEA) and ESEA Flexibility that provide for valid inferences that support the intended uses, such as informing:

- School effectiveness and improvement;
- Individual principal and teacher effectiveness for purposes of evaluation and identification of professional development and support needs;
- Individual student gains and performance; and
- Other purposes defined by the state.

Primary claims related to assessment design:

- A.2.1. The purposes of the assessment, the target population, and each of the intended interpretations and uses of assessment results are clearly articulated.
- A.2.2. The construct or content domain of interest, how it is defined, and the rationale for that specification are clearly articulated.
- A.2.3. The assessment design reflects the construct definition and supports the intended interpretations and uses.
- A.2.4. Documentation is provided that clearly specifies the inferences and assumptions underlying the design of the assessment.

²⁸ The standard setting *methodology* refers to the specific technique or approach used by panelists to recommend performance standards (a.k.a. cut scores) within the context of the standard setting meeting.

Primary claims related to validity evaluation:

- A.2.5. An outline, framework or plan summarizes those studies that have been or will be conducted to collect evidence to support the interpretive argument or validity evaluation plan, including the three primary uses as stated below. (Note: Evidence provided should include both descriptions of planned studies, and documentation /results from completed studies.)
 - A.2.5.a. Evidence is provided to support the use of assessment results for making valid inferences about student performance and readiness for college and career (or on-track to CCR).
 - A.2.5.b. Evidence is provided to support the use of assessment results for making valid inferences about student growth over time.
 - A.2.5.c. Evidence is provided to support the use of assessment results for making valid inferences about school, principal, and teacher effectiveness (if such a use is intended) and informing improvement activities.
- A.2.6. The planned or completed validity evaluation considers the fairness of the assessment program for all examinees with respect to both intended and unintended consequences.
- A.2.7. The design and/or results of planned and/or completed validation studies were reviewed and endorsed by an independent, expert review panel (e.g., technical advisory committee).

While no secondary claims are included directly in the text, criterion A.2 is a special case in that the quality of evidence presented in support of the other criteria will directly influence judgments regarding the validity of score interpretation and use. Therefore, it is essential that when making holistic judgments regarding criterion A.2, consideration be given to the strength of support provided for all other criteria.

A.3 Ensuring that assessments are reliable: Assessments minimize error that may distort interpretations of results, estimate the magnitude of error, and inform users of its magnitude.

Primary claims related to reliability:

- A.3.1. Procedures for quantifying/ calculating reliability indices (e.g. Coefficient alpha, inter-rater reliability, classification accuracy and consistency, generalizability coefficient) and precision (e.g., standard error of measurement with associated confidence bounds, including both overall and conditional SEM, decision-accuracy indices) for each reported score are comprehensive, defensible, and well documented.
- A.3.2. Clear criteria are in place for evaluating the appropriateness of obtained reliability indices and estimates of precision.
- A.3.3. The pre-specified reliability and precision indices were estimated and the results indicate adequate support for intended uses.

Secondary claim related to informing users of reliability: D.1.2

A.4 Ensuring that assessments are designed and implemented to yield valid and consistent test score interpretations within and across years:

- Assessment forms yield consistent score meanings within and across years, as well as for various student groups, and delivery mechanisms (e.g., paper, computer, including multiple computer platforms).
- The score scales facilitate accurate and meaningful inferences about test performance.

Primary claims related to assessment development:

- A.4.1. Item design/ development materials are written at a level of detail that supports appropriate construct coverage and consistency over forms within and across years.
- A.4.2. Items undergo a comprehensive review to ensure they are appropriate, fair, accessible and likely to be interpreted by students in a consistent, accurate manner regardless of group membership or delivery mechanism.
- A.4.3. Item pilot testing and psychometric review procedures are designed to ensure items are fair for all students and provide for valid measures of student performance relative to the construct of interest.
- A.4.4. Test specifications clearly articulate what "equivalence" means from content (KSAs), format and statistical perspectives.
- A.4.5. A comprehensive test review process is in place to ensure test forms meet the content and statistical requirements outlined in the test specifications.

Primary claims related to scaling and equating:

- A.4.6. The design of the scale accounts for the design of the assessment and the manner in which results are intended to be interpreted and used.
- A.4.7. The procedures used to estimate student performance and translate these estimates to a different scale are transparent, fair, and consistent with the reported meaning of the scale scores.
- A.4.8. Procedures for scoring items or sections that involve human judgment (e.g. performance tasks, essays) support accurate and consistent scoring within and across items, forms, administrations, and sub-groups by minimizing construct-irrelevant score variance within and across scorers.
- A.4.9. Linking and/or equating procedures are clearly specified, comprehensive, and demonstratively appropriate.
- A.4.10. The scaling and linking/equating procedures were followed as specified, and the results have been reviewed and accepted by technical experts.

Secondary claims from E.1 related to assessment standardization: E.1.1, E.1.2

A.5 Providing accessibility to all students, including English learners and students with disabilities.

- Assessments provide for reliable scores and valid score interpretations related to intended use for English learners.
- Assessments provide for reliable scores and valid score interpretations related to intended use for students with disabilities.

Primary claims relating to the test characteristics associated with accessibility

- A.5.1. The testing user interface²⁹ and item format does not introduce construct-irrelevant variance that impedes student performance.
- A.5.2. Students are matched with appropriate accommodations/ accessibility features.
- A.5.3. Score reliability is appropriately estimated and evaluated for English learners and students with disabilities (SWD).
- A.5.4. Validity evidence supports the intended use and interpretation of scores for English learners and students with disabilities (SWD).

Secondary claims from A.4 related to item development & review: A.4.2 – A.4.3

A.7 Meeting all requirements for data privacy and access: All assessments must meet federal and state requirements for student privacy, and all data must be readily accessible by the state.

Primary claims related to student privacy:

- A.7.1. Adequate steps have been taken to ensure compliance with Federal Educational Rights and Privacy Act (FERPA) and any additional state regulations related to maintaining student privacy.
- A.7.2. Comprehensive procedures are in place to protect personally identifiable information (PII) from unauthorized access or use.
- A.7.3. Procedures are in place to ensure all data is managed securely.

Primary claims related to data access:

- A.7.4. An assurance is provided of state ownership of all required data³⁰ reflecting compliance with state laws.
- A.7.5. Procedures and timelines are in place to ensure a state³¹ is provided with all data necessary to support desired analyses (e.g., forensics, quality control, accountability calculations) in a timely and useable fashion.
- A.7.6. Procedures are defined for how data will be securely transferred between vendors and the state, and stored or destroyed after administration/reporting.

Secondary claims from E.1 related to security of test materials: E.1.3-E.1.6

²⁹ Testing user interface refers to paper-and-pencil based, computer-based or in some other appropriate item presentation and response-capturing modality.

³⁰ Student performance data includes: student level response strings (scored and unscored), and any associated scores, transformations of scores, and aggregations computed to support reporting.

³¹ When necessary, the state can be replaced to expand to other units that may adopt an assessment such as DODEA, U.S. Territories, large districts in states that choose the local assessment option, and private school organizations.

D.1 Focusing on student achievement and progress to readiness: Score reports illustrate a student's progress on the continuum toward college and career readiness, grade by grade and course by course. Reports stress the most important content skills and processes and how the assessment focuses on them to show whether or not students are on track to readiness.

Primary claims related to score report content and format:

- D.1.1. The content and format of the score reports are consistent with and supported by the assessment design, and the psychometric procedures for developing the scale(s), and support the intended uses.
- D.1.2. Score reports support inferences regarding student achievement relative to key content and performance standards.
- D.1.3. Score reports provide for valid inferences regarding career and college readiness, or on-track to CCR.

D.2 Providing timely data that inform instruction: Reports are instructionally valuable, easy to understand by all audiences and delivered in time to provide useful, actionable data to students, parents and teachers.

Primary claims related to timeliness of score reports:

- D.2.1. Directions for accessing and viewing score reports (when necessary) are broadly distributed and clear to end-users.
- D.2.2. Reporting timelines, procedures and technology provide for the dissemination of test results in a timely fashion.

Primary claims related to instructional utility of score reports:

- D.2.3. The content and structure of score reports provide useful and actionable information for making instructional decisions.

E.1 Maintaining necessary standardization and ensuring test security: in order to ensure the validity, fairness and integrity of state test results, the assessment systems maintain the security of the items and tests as well as the answer documents and related ancillary materials that result from test administration.

Primary claims related to standardization:

- E.1.1. Test distribution and administration directions are clear and sufficiently scripted to provide for standardization.
- E.1.2. Procedures for training and monitoring test administrators are effective and well documented.

Primary claims related to security:

- E.1.3. Comprehensive procedures are in place to ensure the security of assessment materials.
- E.1.4. Effective test security training is provided for all personnel who come into contact with test materials.
- E.1.5. Procedures are in place to test and validate the effectiveness of security safeguards.
- E.1.6. Activities construed as cheating or other breaches of test security are clearly defined and transparent.
- E.1.7. Detailed procedures are in place to support the detection of testing irregularities.
- E.1.8. Clearly documented procedures and specifications are provided for responding to breaches in test security.

APPENDIX C: EXAMPLES OF EVIDENCE THAT MAY BE PROVIDED FOR EVALUATION

Directions:

This appendix is designed to provide examples of the type of evidence (e.g., materials, reports, documents) that may be submitted to support the evaluation of each of the CCSSO criteria. As discussed in The Guide, the specific set of materials provided will vary across tests. The examples outlined below relate specifically to existing assessments. For assessments that are in the proposal phase, or that are still under development, the nature of the evidence provided will clearly differ.

<p>A.1 Indicating progress toward college and career readiness: Scores and performance levels on assessments are mapped to determinations of college and career readiness at the high school level and for other grades being on track to college and career readiness by the time of high school graduation.</p>	<ul style="list-style-type: none"> • State’s definition/interpretation of college-career ready (if applicable) • Summary of PLD development process • PLD technical report, PLD training materials • PLD participant list, including demographics and background information. • PLD participant selection criteria. • Any studies conducted to review or validate the PLDs as reflecting the content standards and CCR expectations. • Standard setting proposal documentation • Final Standard Setting Technical Report – including final recommendations • Standard setting participant list, including demographics and background information • SS panelist training materials. • Description of studies planned to support validation of the performance standards. • Any written reports or independent evaluations of the standards setting process which shows it was conducted as intended. • Evidence showing that standard setting materials and results were reviewed and approved by an external technical committee • Standard setting panelist survey results. • Documentation reflecting the participation of experts in the development of the PLD and/or Standard setting design process – (e.g., TAC summaries, feedback from reviewers, etc...)
<p>A.2 Ensuring that assessments are valid for required and intended purposes: Assessments produce student achievement and student growth data, as required under Title 1 of the Elementary and Secondary Education Act (ESEA) and ESEA Flexibility, that provide for valid inferences that support the intended uses, such as informing:</p>	<ul style="list-style-type: none"> • Technical manual • Theory of Action/Interpretive Argument/Assessment Argument/Design Rationale • Description of the assessment domain and how it was determined. • Documentation of test development process/rationale • Description of and rationale for procedures used to measure student growth • Summary of implemented and planned validity studies • Any studies collected to support the use and interpretation of results as intended (especially for making CCR or on-track inferences) • Documentation showing the involvement of experts in the review, design and/or planning of validation studies.

<p>A.3 Ensuring that assessments are reliable: Assessments minimize error that may distort interpretations of results, estimate the magnitude of error, and inform users of its magnitude.</p>	<ul style="list-style-type: none"> • Data analysis procedures related to reliability. • Statistical test development specifications including statistical targets for test form difficulty or other relevant objective functions and reliability evaluation criteria (test information functions or reliability coefficients) • Technical manual – including results of any conducted reliability analyses (e.g., overall, by sub-group, by reportable category) • Research studies or analyses conducted to evaluate reliability • Documentation showing the involvement of experts in the review of reliability outcomes
<p>A4. Ensuring that assessments are designed and implemented to yield valid and consistent test score interpretations within and across years:</p> <ul style="list-style-type: none"> • Assessment forms yield consistent score meanings within and across years, as well as for various student groups, and delivery mechanisms (e.g., paper, computer, including multiple computer platforms). <p>The score scales facilitate accurate and meaningful inferences about test performance.</p>	<ul style="list-style-type: none"> • Item development specifications (including eligible content, evidence statements and format/precision specifications) • Item writer training materials • Summary of item review, revision, and approval process • Evidence showing the participation of Universal Design experts in the development/review process • Test Development Procedures and specifications • Scoring procedures and rationale • Scaling and equating procedures and rationale • Evidence of expert review of equating procedures/results. • Field-Testing and data review procedures and summaries of results.
<p>A5. A.5 Providing accessibility to all students, including English learners and students with disabilities.</p> <ul style="list-style-type: none"> • Following the principles of universal design • Offering appropriate accommodations and modifications: <ul style="list-style-type: none"> • Assessments provide for reliable scores and valid score interpretations related to intended use for English learners. • Assessments provide for reliable scores and valid score interpretations related to intended use for students with disabilities. 	<ul style="list-style-type: none"> • Results of assessment usability studies and/or cognitive labs • Planned or conducted research related to accessibility • Procedures/guidelines developed to support the selection of appropriate student accommodations • Rules/guidelines related to the administration of practice tests prior to operational administration. • Guidelines to support test interpreters, when necessary • Reports or summaries of feedback from educators around procedures used to assign/administer accommodated forms. • Technical report – including any reliability analyses conducted for EL and/or SWD. • Summaries of studies planned or implemented to verify validity of score-based inferences for all students.
<p>A.7 Meeting all requirements for data privacy and access: All assessments must meet federal and state requirements for student privacy, and all data must be readily accessible by the state</p>	<ul style="list-style-type: none"> • Data security procedures • Data protection protocols • Security breach procedures • Procedures, specifications related to the ownership of data during and after contract. • Procedures for ensuring secure data transfer • Procedures and specifications for destroying secure data
<p>D.1 Focusing on student achievement and progress to readiness: Score reports illustrate a student’s progress on the continuum toward college and career readiness, grade by grade and course by course. Reports stress the most important content skills and processes and how the assessment focuses on them to show whether or not students are on track to readiness.</p>	<ul style="list-style-type: none"> • Summary of process used to develop, evaluate and revise score reports prior to operational use • Samples of score reports, interpretive guides • Evidence of focus groups and usability analyses conducted to support the development of score reports and interpretive materials. • Documentation outlining the purpose and intended user of each report.

<p>D.2 Providing timely data that inform instruction: Reports are instructionally valuable, easy to understand by all audiences and delivered in time to provide useful, actionable data to students, parents and teachers.</p>	<ul style="list-style-type: none"> • Score report timelines and plans for distribution • Summary of process used to develop, evaluate and revise score reports prior to operational use • Samples of score reports, interpretive guides, results of focus groups • Directions provided to support access to and viewing of score reports
<p>E.1 Maintaining necessary standardization and ensuring test security: in order to ensure the validity, fairness and integrity of state test results, the assessment systems maintain the security of the items and tests as well as the answer documents and related ancillary materials that result from test administration.</p>	<ul style="list-style-type: none"> • Test administration guidelines • Directions to proctors and test takers • Data Management and Security Procedures • Descriptions of security procedures related to: item/test development, review, publishing and administration; exposure control procedures for CAT. • Evidence or documentation that directions provided in operational forms and test administration guidelines are clear (e.g., usability analyses, feedback reports, internal review procedures.)

APPENDIX D: GUIDELINES AND TEMPLATES TO SUPPORT THE COLLECTION AND ORGANIZATION OF EVIDENCE

This appendix is designed to provide guidelines to support the identification and organization of evidence, to support evaluation. Those charged with assembling the evidence provided for review are responsible for determining what (and how much) evidence is necessary and appropriate to support the evaluation of each claim. In doing so they should strive to identify *the minimum amount of evidence necessary to allow for qualified technical experts to make informed and reliable determinations regarding the extent to which a given claim is supported*. Some general guidelines to support this process are provided below:

COMMENTS AND GUIDELINES TO SUPPORT THE IDENTIFICATION AND ORGANIZATION OF EVIDENCE

- Use the text of the claim and associated sufficiency statements as a guide in determining what is/is not necessary to support evaluation. Information that is interesting (and only marginally related), but not necessary should be considered extraneous.
- Put yourself in the evaluator's shoes. What information do they need to understand and interpret the evidence provided? For example, are there program-specific terminologies or acronyms that require definition?
- If the same *procedures, methods and statistical criteria* are generally used across multiple grades and content areas (e.g., those related to test development/review, scaling and equating procedures, standard setting, etc....) it is not necessary to provide documentation of these procedures (and the accuracy with which they were implemented) multiple times. A limited sample of evidence can be provided with a comment around the grades/content areas to which it generalizes.
- If, there are *important* procedural differences across grades/subjects that are *relevant to the evaluation of a given claim* (e.g., differences in the type of external data provided to support standard setting at the high-school vs the elementary school level) the information and context necessary to support evaluation of both procedures should be provided.
- When claims *require the review of key outputs that would vary across grades/subjects*, (e.g., validity evidence supporting the CCR and on-track to CCR inferences; reliability coefficients; score reports) relevant results and documentation should be provided for all grades/content areas.
- If a claim is asking about the manner in which a particular type of data is represented, described, formatted or presented; exemplary samples of that output, rather than all instances should be sufficient to support evaluation.
- Documentation related to the endorsement of evidence by an external review committee should be detailed enough so that evaluators will know what the committee reviewed, the nature of the discussion surrounding the relevant material and the recommendations that resulted. An agenda for a meeting does not meet this requirement.
- To assure accuracy and efficiency in the evaluation process, in the narratives directing evaluators to relevant evidence for any given claim, evaluators should be provided with clear document references, page numbers, and paragraph/table/figure citations to guide them directly to the relevant evidence. The documents themselves should be marked up to highlight the relevant evidence (e.g., drawing red boxes around the relevant evidence, highlighting in yellow, annotating which claim(s) the marked up content addresses). If the context of an entire document is not deemed relevant, an excerpt may be provided for brevity, but a full reference to the document should be provided.
- To ensure access to the evidence, all documents should be provided in commonly used formats, such as PDF and Microsoft Office.

It is assumed that, for most tests, there will be at least *some* evidence provided to support each claim. If a specific type of evidence is unavailable or believed not to be applicable given the goals of the test, the Provider should include comments explaining why this is the case.

When collecting and organization evidence it is extremely important that those charged with organizing the set of materials that will be submitted to the Implementation Team for evaluation (i.e., the Provider) note the following:

1. Providers will not be given multiple opportunities to provide evidence to the Implementation Team once the evaluation process has begun (i.e., the Evaluation Team has been provided with the materials for review). Therefore the Provider must meet all requirements for evidence the *first time around*. The provision of data/evidence is not an iterative process!

2. There are several places in the criteria evaluation framework (CEF) where process-based documentation must be supplemented by evidence that the process actually occurred. Providers should read the CEF carefully to ensure that, when necessary, evidence that a process occurred as intended is provided to the evaluation team for review (e.g., outcomes of technical advisory meetings, information gained through external review of procedures/materials, etc.).
3. Evaluators will be looking for a coherent validity argument³² in the evidence provided. They are not planning to have to construct one on their own! When multiple pieces of evidence must be considered simultaneously in support of a given claim or criterion, the Provider is responsible for reflecting this in their submission. It should not be assumed that the evaluator will put the pieces together as needed to support the evaluation.

SAMPLE EVIDENCE LOG

In the table below, you will see an example of an evidence log that includes reports that could be used in the evaluation of a hypothetical assessment program. This document is to be completed by the Providers. This example is provided to help the organizations submitting their own documents and to illustrate the type of descriptions that would be appropriate.

DOCUMENT NUMBER	DOCUMENT NAME	BRIEF DESCRIPTION
1	State Technical Report (2013-2014)	Technical Report – includes grades 3-8 Math, Reading, Writing and Science
2	PLD participant lists	PLD Participant summaries, including information about background and experience
3	Sample of Score Reports	Sample student, school, roster, and teacher reports
4	Score Report Interpretive Guides	Materials provided to support test users in interpreting each score report
5	TAC Materials and Meeting Minutes associated with Standard Setting Review	Minutes summarizing the nature of the TAC review of SS results, materials provided for review and any TAC comments.
6		
7		
.		

When submitting your evidence log, for each piece of evidence submitted, enter the document name (as it appears in the directory or electronic evidence file) and provide a brief description of its contents. If there is a set of interdependent documents that will always be viewed together, these can be concatenated and assigned one document number.

³² An argument or rationale which describes how the evidence provided lends support to the overall claim or criterion. This is especially important for A.2.

PRIORITIZED EVIDENCE LIST

In the table below, you will see an example of a prioritized evidence list that includes reports that could be used in the evaluation of a hypothetical assessment program. This document is intended to be completed by the Provider. This example is provided to help the organizations submitting documentation and to illustrate the type of descriptions that would be appropriate.

Each piece of evidence should be referenced by the document number assigned on the E-log and, when appropriate, those page numbers (chapters, appendices, etc.) most relevant to evaluating the claim should be noted. If the relevance of a particular piece of evidence is not readily apparent or additional background information is necessary to support its review, this should be noted in the comments section. Similarly, if two pieces of evidence are linked, or should be jointly considered in service to a given claim, this should also be noted with comments. It is assumed that, for most tests, there will be at least some evidence provided to support each claim. If a specific type of evidence is unavailable or believed to be not applicable given the goals of the test, the provider should include comments explaining why this is the case.

CLAIM	CLAIM	MATERIALS TO BE REVIEWED	COMMENTS
A1.1	College- and career readiness has been clearly defined for operational use.	#1, Pages 3-5	Operational definition of "College- and career-ready" including references to relevant research to support the definition
A1.2	The process for developing performance level descriptors (PLDs) provides for PLDs that accurately represent the expectations defined by the CCR content standards within and across grades.	#1, Pages 25-40	Provides a comprehensive summary of the PLD development process and results
		#2, Participant List	This list reflects participants at the initial set of PLD meetings.
		#, TAC materials, Chapter 3	Additional experts involved in the evaluation and finalization of PLDs are provided in Document #5, the materials from the standard setting TAC meeting.

When submitting your evidence summary, for each evidence statement, enter the document number and name (as it appears in the directory or electronic evidence file) and a brief description of its contents. When supplying this information, provide as much specificity as possible, whether it is specific chapters in a technical manual, or specific pages from the minutes of a TAC meeting. In the comments section, information should be supplied to help the evaluators fully comprehend how the document supports the evidence statement.

APPENDIX E: INDEPENDENT EVALUATOR RATING SHEET SAMPLE

INDIVIDUAL EVALUATOR RATING SHEET FOR CRITERION A.1

For each claim make a holistic rating of the degree to which the body of evidence provided lends support to that claim. Circle this rating in the left-hand column of the rating sheet provided for that claim. After making your rating, briefly describe your rationale for the rating focusing on those factors/pieces of evidence you found most influential. If any of the evidence submitted for a given claim violates your expectations, clearly articulate this fact within the provided rationale.

RATING	
Does Not Meet	There is either no evidence to support the given claim, or, the evidence provided does not lend support to the claim statement as written. The evidence provided may not be relevant, or is extremely insufficient given the assessment's current phase of development or implementation.
Partially Meets	The evidence reviewed provides some support for the claim, and while the claim is partially supported, the evaluator has reservations about endorsing the claim fully. The evidence provided may be incomplete, or not as comprehensive as expected given the assessment's current phase of development or implementation.
Meets	The evidence reviewed provides clear and unequivocal support for the claim as written. The evidence provided is comprehensive and appropriate given the assessment's current phase of development or implementation.

A.1 Indicating progress toward college and career readiness: Scores and performance levels on assessments are mapped to determinations of college and career readiness at the high school level and for other grades being on track to college and career readiness by the time of high school graduation.

A.1.1 College- and career readiness has been clearly defined for operational use.	
To what extent is this claim supported given the evidence provided? <i>Does not meet</i> <i>Partially meets</i> <i>Meets</i>	Comments:
Claim: A.1.2 The process for developing performance level descriptors (PLDs) provides for PLDs that accurately represent the expectations defined by the CCR content standards within and across grades.	
To what extent is this claim supported given the evidence provided? <i>Does not meet</i> <i>Partially meets</i> <i>Meets</i>	Comments:

Claim: A.1.3 Knowledgeable experts were involved in the process of developing and reviewing the PLDs	
To what extent is this claim supported given the evidence provided? <i>Does not meet</i> <i>Partially meets</i> <i>Meets</i>	Comments:
Claim: A.1.4 The process used for developing performance level descriptors (PLDs) supports their intended use(s).	
To what extent is this claim supported given the evidence provided? <i>Does not meet</i> <i>Partially meets</i> <i>Meets</i>	Comments:
Claim: A.1.5 The process for developing performance level descriptors (PLDs) includes an evaluation of alignment of the PLDs to the content of the test questions that differentiate performance at each level, and, as needed, re-writing based on new evidence concerning skills needed for success in college and careers.	
To what extent is this claim supported given the evidence provided? <i>Does not meet</i> <i>Partially meets</i> <i>Meets</i>	Comments:
Claim: A.1.6 A description and coherent rationale are provided for how the proposed and/or implemented standard setting methodology³³ yields valid determinations of progress toward, or attainment of, college and career readiness	
To what extent is this claim supported given the evidence provided? <i>Does not meet</i> <i>Partially meets</i> <i>Meets</i>	Comments:

³³ The standard setting *methodology* refers to the specific technique or approach used by panelists to recommended performance standards (a.k.a. cut scores) within the context of the standard setting meeting.

Claim: A.1.7 A coherent rationale accompanies methodological decisions regarding the level of involvement of grade-level educators, higher education, industry, and career technical experts (CTEs) in the standard setting process.	
To what extent is this claim supported given the evidence provided? <i>Does not meet</i> <i>Partially meets</i> <i>Meets</i>	Comments:
Claim: A.1.8 Appropriate external CCR benchmarks and research studies are/were used in the standard setting process.	
To what extent is this claim supported given the evidence provided? <i>Does not meet</i> <i>Partially meets</i> <i>Meets</i>	Comments:
Claim: A.1.9 Procedures and rationales for any adjustments made to proposed cut scores after the standard setting meeting are based on a defensible rationale and method.	
To what extent is this claim supported given the evidence provided? <i>Does not meet</i> <i>Partially meets</i> <i>Meets</i>	Comments:
Claim: A.1.10 Studies planned or conducted to evaluate the validity of CCR performance standards over time are appropriate given the inferences they are intended to support.	
To what extent is this claim supported given the evidence provided? <i>Does not meet</i> <i>Partially meets</i> <i>Meets</i>	Comments:

Claim: A.1.11

The standard setting procedures were followed as specified, and the final cut scores and the results of validity studies have been reviewed by technical experts.

To what extent is this claim supported given the evidence provided?

Does not meet

Partially meets

Meets

Comments:

Comments Related to Secondary Claims A4.6-A4.10

OVERALL RATING FOR CRITERIA A.1

WEAK	LIMITED	GOOD	EXCELLENT
<p>The body of evidence presented provides weak support for this criterion, as defined in the Criteria Evaluation Framework. The evidence reviewed consistently did not meet expectations given the assessment’s current phase of development or implementation, design, and specified purpose(s) and goals.</p>	<p>The body of evidence presented provides limited support for this criterion, as defined in the Criteria Evaluation Framework. While some of the evidence may have met expectations, key elements central to this criterion are missing or poor given the assessment’s current phase of development or implementation, design, and specified purpose(s) and goals.</p>	<p>The body of evidence presented provides adequate support for this criterion as defined in the Criteria Evaluation Framework. The preponderance of evidence reviewed met expectations given the assessment’s current phase of development or implementation, design, and specified purpose(s) and goals.</p>	<p>The body of evidence presented provides strong support for this criterion, as defined in the Criteria Evaluation Framework. The evidence reviewed consistently met or exceeded expectations given the assessment’s current phase of development or implementation, design, and specified purpose(s) and goals. The evidence could serve as a strong exemplar or model for similar assessments.</p>

Program-specific narrative guide for interpreting overall rating of criterion:

APPENDIX F: INDEPENDENT EVALUATOR RATING SHEET SAMPLE SPECIFIC TO CRITERION A.5

INDIVIDUAL EVALUATOR RATING SHEET FOR CRITERION A.5

For each claim make a holistic rating of the degree to which the body of evidence provided lends support to that claim. Circle this rating in the left-hand column of the rating sheet provided for that claim. After making your rating, briefly describe your rationale for the rating focusing on those factors/pieces of evidence you found most influential. If any of the evidence submitted for a given claim violates your expectations, clearly articulate this fact within the provided rationale.

Evidence presented regarding students with disabilities and English learners may be considered separately. Ratings and comments relative to these student groups can be captured independent from one another, where indicated on this sample rating sheet.

RATING	
Does Not Meet	There is either no evidence to support the given claim, or, the evidence provided does not lend support to the claim statement as written. The evidence provided may not be relevant, or is extremely insufficient given the assessment's current phase of development or implementation.
Partially Meets	The evidence reviewed provides some support for the claim, and while the claim is partially supported, the evaluator has reservations about endorsing the claim fully. The evidence provided may be incomplete, or not as comprehensive as expected given the assessment's current phase of development or implementation.
Meets	The evidence reviewed provides clear and unequivocal support for the claim as written. The evidence provided is comprehensive and appropriate given the assessment's current phase of development or implementation.

A.5 Providing accessibility to all students, including English learners and students with disabilities.

- Assessments provide for reliable scores and valid score interpretations related to intended use for English learners.
- Assessments provide for reliable scores and valid score interpretations related to intended use for students with disabilities.

Claim: A.5.1

The testing user interface and item format does not introduce construct-irrelevant variance that impedes student performance.

To what extent is this claim supported given the evidence provided?

Does not meet
Partially meets
Meets

Comments:

Claim: A.5.2

Students are matched with appropriate accommodations/ accessibility features.

To what extent is this claim supported given the evidence provided?

Does not meet
Partially meets
Meets

Comments:

Claim: A.5.3

Score reliability is appropriately estimated and evaluated for English learners and students with disabilities (SWD).

To what extent is this claim supported given the evidence provided?

Does not meet

Partially meets

Meets

Comments:

To what extent is this claim supported given the evidence provided relative to students with disabilities?

Does not meet

Partially meets

Meets

Comments:

Claim: A.5.4

Validity evidence supports the intended use and interpretation of scores for English learners and students with disabilities (SWD).

To what extent is this claim supported given the evidence provided?

Does not meet

Partially meets

Meets

Comments:

To what extent is this claim supported given the evidence provided relative to students with disabilities?

Does not meet

Partially meets

Meets

Comments:





















OVERALL RATING FOR CRITERIA A.5













WEAK	LIMITED	GOOD	EXCELLENT
<p>The body of evidence presented provides weak support for this criterion, as defined in the Criteria Evaluation Framework. The evidence reviewed consistently did not meet expectations given the assessment’s current phase of development or implementation, design, and specified purpose(s) and goals.</p>	<p>The body of evidence presented provides limited support for this criterion, as defined in the Criteria Evaluation Framework. While some of the evidence may have met expectations, key elements central to this criterion are missing or poor given the assessment’s current phase of development or implementation, design, and specified purpose(s) and goals.</p>	<p>The body of evidence presented provides adequate support for this criterion as defined in the Criteria Evaluation Framework. The preponderance of evidence reviewed met expectations given the assessment’s current phase of development or implementation, design, and specified purpose(s) and goals.</p>	<p>The body of evidence presented provides strong support for this criterion, as defined in the Criteria Evaluation Framework. The evidence reviewed consistently met or exceeded expectations given the assessment’s current phase of development or implementation, design, and specified purpose(s) and goals. The evidence could serve as a strong exemplar or model for similar assessments.</p>

Program-specific narrative guide for interpreting overall rating of criterion:

APPENDIX G: SAMPLE TEST CHARACTERISTICS SUMMARY REPORT TEMPLATE

Assessment Phase of Development: _____

	Degree of Match with CCSSO Criteria			
	Weak	Limited	Good	Excellent
<p>A.1 Indicating progress toward college and career readiness: Scores and performance levels on assessments are mapped to determinations of college and career readiness at the high school level and for other grades to being on track to college and career readiness by the time of high school graduation.</p> <p><i>[[summary of rationale and other comments]]</i></p>				
<p>A.2 Ensuring that assessments are valid for required and intended purposes: Assessments produce data, including student achievement data and student growth data required under Title I of the Elementary and Secondary Education Act (ESEA) and ESEA Flexibility, that can be used to validly inform the following:</p> <ul style="list-style-type: none"> • School effectiveness and improvement; • Individual principal and teacher effectiveness for purposes of evaluation and identification of professional development and support needs; • Individual student gains and performance; and • Other purposes defined by the state. <p><i>[[summary of rationale and other comments]]</i></p>				
<p>A.3 Ensuring that assessments are reliable: Assessments minimize error that may distort interpretations of results, estimate the magnitude of error, and inform users of its magnitude.</p> <p><i>[[summary of rationale and other comments]]</i></p>				
<p>A.4 Ensuring that assessments are designed and implemented to yield valid and consistent test score interpretations within and across years:</p> <ul style="list-style-type: none"> • Assessment forms yield consistent score meanings within and across years, as well as for various student groups, and delivery mechanisms (e.g., paper, computer, including multiple computer platforms). • The score scales facilitate accurate and meaningful inferences about test performance. <p><i>[[summary of rationale and other comments]]</i></p>				
<p>A.5 Providing accessibility to all students, including English learners and students with disabilities:</p> <ul style="list-style-type: none"> • Assessments produce valid and reliable scores for English learners • Assessments produce valid and reliable scores for students with disabilities. <p><i>[[summary of rationale and other comments]]</i></p>				

	Degree of Match with CCSO Criteria			
	Weak	Limited	Good	Excellent
<p>A.7 Meeting all requirements for data privacy and ownership: All assessments must meet federal and state requirements for student privacy, and all data is owned exclusively by the state.</p> <p><i>[[summary of rationale and other comments]]</i></p>				
<p>D.1 Focusing on student achievement and progress to readiness: Score reports illustrate a student's progress on the continuum toward college and career readiness, grade by grade, and course by course. Reports stress the most important content, skills, and processes, and how the assessment focuses on them, to show whether or not students are on track to readiness.</p> <p><i>[[summary of rationale and other comments]]</i></p>				
<p>D.2 Providing timely data that inform instruction: Reports are instructionally valuable, easy to understand by all audiences and delivered in time to provide useful, actionable data to students, parents and teachers.</p> <p><i>[[summary of rationale and other comments]]</i></p>				
<p>E.1 Maintaining necessary standardization and ensuring test security: in order to ensure the validity, fairness and integrity of state test results, the assessment systems maintain the security of the items and tests as well as the answer documents and related ancillary materials that result from test administration.</p> <p><i>[[summary of rationale and other comments]]</i></p>	