The Role of the *Standards for Educational and Psychological Testing* in Establishing a Methodology to Support the Evaluation of Assessment Quality[1]

Susan Lyons, Ph.D.
Erika Hall, Ph.D.

Introduction

The *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 2014) is regarded as one of the primary reference documents for assessment and measurement professionals. It defines the criteria against which test materials, results and practices must be evaluated and clarifies who is responsible for ensuring that certain expectations defined within the *Standards* are met (i.e., test sponsor, developer, publisher, administrator, users, etc.). While consideration of the *Standards* is often second-nature to test development vendors, assessment consultants and academics in the cognitive/behavioral sciences, the tendency to attend to the criteria outlined in the *Standards* often does not generalize to those selecting, evaluating and using assessments within state, district or classroom settings. This is likely true for a variety of reasons including: a general lack of familiarity with the *Standards* and their purposes, which is often reflected by the perception that the *Standards* are only important for/useful to test vendors; lack of understanding of many of the assessment concepts necessary to interpret and use the *Standards*; difficulty navigating and applying the *Standards* in service to a given purpose or use; and lack of clarity around one's role(s) in the testing process and, consequently, who is responsible for what when it comes to adherence to the *Standards*.

[1] Paper presented at the 2016 Annual Meeting of the National Council for Measurement in Education (NCME) in Washington, DC

For example, a superintendent charged with selecting an assessment for use as a measure of student growth within his/her district's educator evaluation system may look to the *Standards* to inform decisions regarding the appropriateness of one or more tests for this purpose. Upon doing so, however, the superintendent may have difficulty identifying those standards the district is responsible for addressing versus those which fall to the vendor, because the chapters are not organized to support evaluation relative to any particular test use. This is no fault of the *Standards*, as the document could never fully address all the different contexts in which tests might be developed and used, but this example simply acknowledges the fact that a certain level of "standards literacy" is necessary if we desire broader use of the document as a guide to "support the development and evaluation of tests and testing practices, and provide guidelines for assessing the validity of interpretations of test scores for the intended test uses" (AERA, APA & NCME, 2014, p. 1) by those without formal training in assessment or measurement. Given the recent policy shift to provide more local control related to the selection and use of assessments, as reflected in certain provisions of the *Every Student Succeeds Act*, the need for training focused on how and when the *Standards* should be used is crucial. If school, districts, and state leaders take on a larger role in identifying assessments that contribute to accountability, it will fall to them to defend the tools and associated measures selected for this use.

Unfortunately, even if training related to use of the *Standards* is developed, the document will still be a heavy lift for many given its breadth and complexity and the degree of professional judgment necessary to determine which *Standards* are more or less relevant in a particular context. Therefore, we believe it is also important that any tools developed to support those charged with making decisions about tests and how they are used be explicitly linked to the *Standards* so the relevance of the document and the manner in which different standards are

addressed within a given context is clear. Specifically, test development, design, analysis and validation procedures and reports generated by assessment vendors, consultants or others supplying materials to test consumers should clearly indicate adherence to best practice through a transparent link to the *Standards*. In addition, to the extent possible, when there is a standard or expectation which falls specifically to the test user to address, this fact should be made explicit in provided documentation.

The purpose of this document is to describe the manner in which the *Standards* were used by the Center for Assessment to inform the development of a framework for evaluating assessments against CCSSO's *Criteria for Procuring and Evaluating High Quality Assessments*. We begin by providing a brief introduction to the CCSSO assessment quality criteria and the Criteria Evaluation Framework. The Criteria Evaluation Framework was developed by the Center to support the evaluation of evidence related to those CCSSO criteria reflecting the technical properties (e.g., reliability and validity) and administration-related concerns (e.g., test security and score reporting) of a given assessment. Next, we discuss the process used to ensure adequate and appropriate representation of the *Standards* given the format and structure of the evaluation framework, and highlight those areas where it was necessary to go outside of the *Standards* to find information regarding best practices. Additionally, we discuss the added value and utility we believe the Criteria Evaluation Framework brings to the *Standards* for the particular use case for which it was developed. Finally, we discuss the benefits of organizing the evaluation framework in terms of claims and evidence statements and describe how this general structure provides a way to establish a coherent link between the *Standards* and assessment practices which require their consideration and adherence.

Background

The *Criteria for Procuring and Evaluating High Quality Assessments* (a.k.a., the Criteria) were developed by the Council of Chief State School Officers (CCSSO) to support states as they "develop procurements and evaluate options for high-quality state summative assessments aligned to their college- and career readiness *Standards*" (CCSSO, 2014, p. 1). The CCSSO criteria are grouped into five broad categories:

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

Recently, at the request of the High Quality Assessment Project[2], the Center for Assessment developed methodologies and procedures to help guide persons and organizations interested in evaluating assessments against CCSSO's criteria. To facilitate this process, Center for Assessment researchers grouped the 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 content 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. A table

[2] 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.

summarizing the specific CCSSO Criteria addressed by each methodology is provided in Figure 1.

TEST CONTENT	TEST CHARACTERISTICS
A. Meet Overall Assessment goals and Ensure Technical Quality - A.5 Providing accessibility to all students, including English learners and students with disabilities (partial) - A.6.Ensuring transparency of test design expectations. **B. Align to Standards - English Language Arts/Literacy** - B.1 Assessing student reading and writing achievement in both ELA and literacy - B.2 Focusing on complexity of texts - B.3 Requiring students to read closely and use evidence from texts - B.4 Requiring 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 **C. Align to Standards - Mathematics** - C.1 Focusing strongly on the content most needed for success in later mathematics - C.2 Assessing a balance of concepts, procedures, and applications - C.3 Connecting practice to content - C.4 Requiring a range of cognitive demand - C.5 Ensuring high-quality items and a variety of item types	**A. Meet Overall Assessment goals and Ensure Technical Quality** - A.1 Indicating progress toward college and career readiness - A.2 Ensuring that assessments are valid 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 **D. Yield Valuable Reports on Student Progress and Performance** - D.1 Focusing on student achievement and progress to readiness - D.2 Providing timely data that inform instruction **E. Adhere to Best Practices in Test Administration** - E.1 Maintaining necessary standardization and ensuring test security

Figure 1. Organization of CCSSO Criteria into Test Content and Test Characteristics

While evaluation procedures were developed for both sets of criteria, this paper focuses on materials developed to support the test characteristics methodology and associated framework, which addresses those criteria reflected in the right-hand column of Figure 1.

Brief overview of the Criteria Evaluation Framework[3]

The test characteristics evaluation methodology and the associated Criteria Evaluation Framework were designed specifically to support the evaluation of high-stakes summative assessments developed to meet the accountability requirements of the No Child Left Behind act of 2001 (NCLB) and/or ESEA waivers. For this reason, it was understood from the onset that the body of evidence and the expertise necessary to implement a comprehensive assessment

[3] Both the test characteristics evaluation methodology and the associated Criteria Evaluation framework can be found at the Center for Assessment's website: http://www.nciea.org/aqem-resources/

evaluation would be significant. It is important to note that the test characteristics and test content methodologies were developed independent from the U.S. Department of Education's Peer Review Guidance updated in September of 2015. However, the evaluation reports generated as a result of applying these methodologies could provide important evidence for use in state submissions to peer review.

While the original CCSSO Criteria document provides a strong foundation upon which to build an evaluation of assessment quality, additional detail and structure were necessary to support the specification of a coherent evaluation process. To this end, the Center developed an evaluation tool referred to as the Criteria Evaluation Framework. The Criteria Evaluation Framework 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).

While a detailed description of the Criteria Evaluation Framework is outside the scope of this document, Figure 2 illustrates the hierarchical structure of the framework and its component parts.

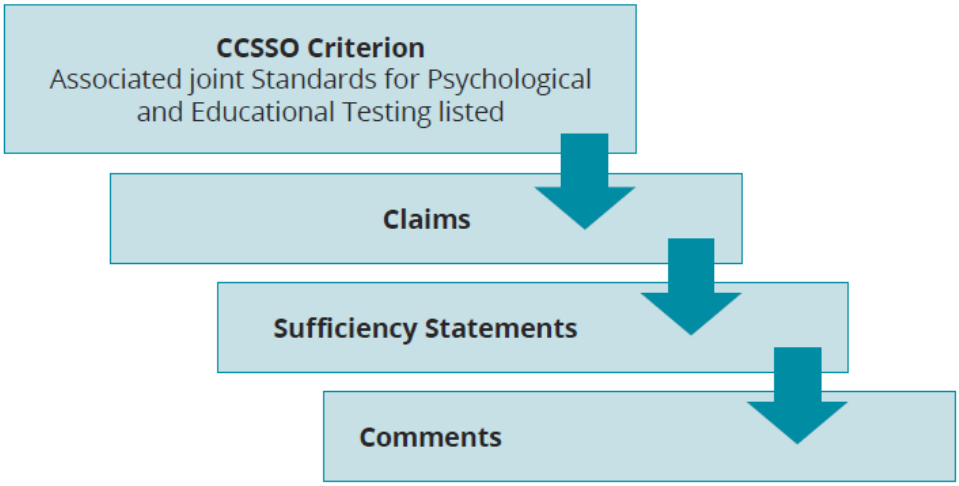


Figure 2. Hierarchical Structure of the Criteria Evaluation Framework

As shown in Figure 2, for each criterion, there is one or more associated claims. *Claims are statements we want to make about procedures, materials, reports, and/or data given the evidence provided for review.* 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. 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 high quality evidence should or may look like and factors that could influence the manner in which that evidence is evaluated 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. Those involved in conducting the evaluation will be asked to consider the features described in the sufficiency statements in addition to contextual factors, such as the assessment's current phase of development, to determine the degree to which each claim and criterion are

supported. To inform this process, comments and examples are provided to highlight how contextual factors may influence one's thinking about the quality of evidence submitted for a given test. *Comments are included as additional notes to aid reviewers in judging the quality of evidence within the context of an assessment program.*

To illustrate how this hierarchical structure is represented in the Criteria Evaluation Framework, one of the six claims associated with Criterion A.7, reflecting requirements for data privacy and access, is provided in Appendix A.

The Criteria Evaluation Framework is the primary tool used by evaluators to support assessment evaluation within the context of the test characteristics evaluation methodology. The methodology includes four phases including an independent review and evaluation of evidence by each evaluator followed by team discussion and the development of a consensus opinion regarding the extent to which provided evidence lends support to each criterion. An overview of each phase is provided in Figure 3.

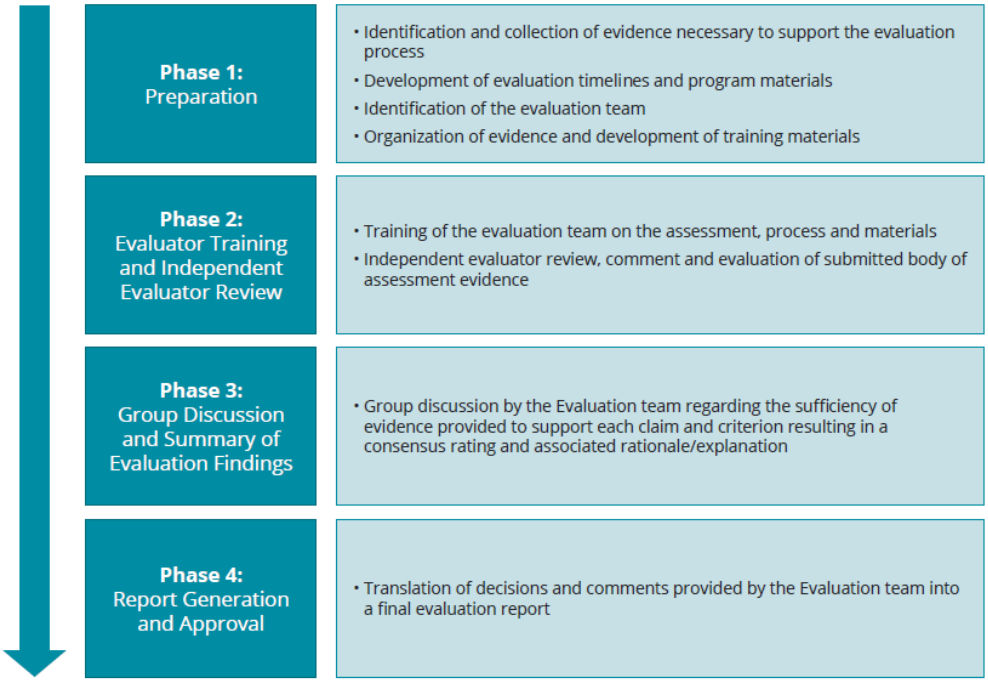


Figure 3. Phases of the Test Characteristics Evaluation Methodology

In the sections which follow we discuss the role of the *Standards* in developing the Criteria Evaluation Framework. The reader is directed to the complete Evaluation Methodology on the Center's website[4] for additional detail regarding the process for conducting a comprehensive evaluation using these materials.

Use of the *Standards* in defining the Criteria Evaluation Framework

When developing and reviewing the Criteria Evaluation Framework, we relied heavily on the *Standards for Educational and Psychological Testing* (2014). When writing the sets of claims and sufficiency statements that structured the evaluation framework for each of the criteria, we used to *Standards* to supplement and validate our initial thinking. Additionally, we used the *Standards* to review the Criteria Evaluation Framework to ensure that the definition of high quality evidence, as described by the sufficiency statements, maintained the expectations of assessments programs as outlined in the *Standards*.

Throughout the development of the methodology and the Criteria Evaluation Framework, the materials associated with the effort underwent a series of internal and external reviews, including the implementation of a test case. Figure 3, below, broadly summarizes the steps in the development process.

[4] http://www.nciea.org/aqem-resources/

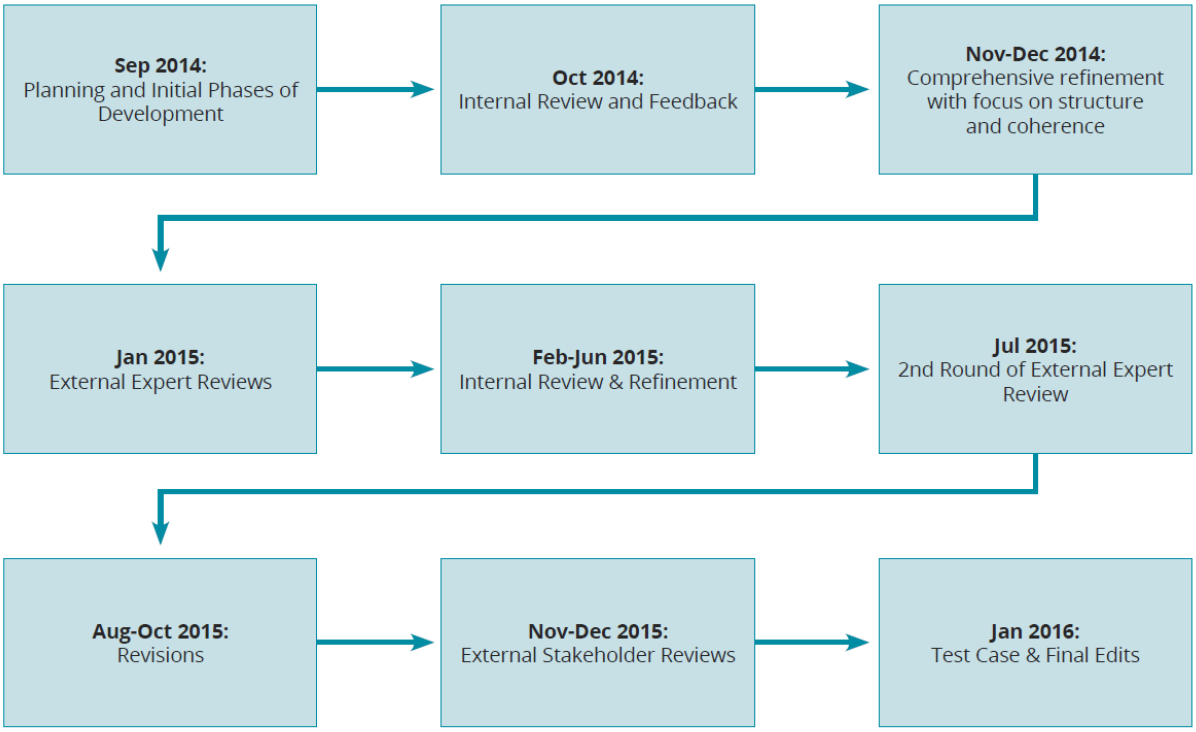


Figure 3. Development Timeline

In September 2014, during the project planning and initial phases of development, associates at the Center for Assessment worked individually, taking the lead on different criteria, to begin to develop the claims and language of the sufficiency statements. At this point in the process, we used the *Standards* to ensure full coverage of criteria, and to supplement the initial thinking of the Center staff. For example, when developing the language to support Criterion D.1 regarding score reporting, we took guidance from standards 1.3, 1.14, 1.15, 5.1, 5.2, 6.10, 12.11, 12.18. These standards are taken from the chapters on Validity, Interpretations of Scores, Reporting and Interpretation, and Educational Testing and Assessment. To further illustrate this example, Table 1 shows the link between standards 1.3, 5.1 and 6.10 and much of the language in the sufficiency statement for Claim D.1.2.[5]

[5] Claim D.1.2 reads "Score reports support inferences regarding student achievement relative to key content and performance standards."

Table 1
Evidence of the Standards in Criteria Evaluation Framework

Standards	Evidence of Standards in Sufficiency Statement and Comments
1.3: "If validity for some common of likely interpretation for a given use has not been evaluated, or if such an interpretation is inconsistent with available evidence, that fact should be made clear and potential users should be strongly cautioned about making unsupported interpretations" (p. 23).	Documentation is provided that indicates the intent of each score report and its primary audience. … The intent of each score report is clearly specified as is the manner in which each reported score is to be interpreted and used (in general and in light of reliability/precision data as well as pertinent validity evidence). The data/information presented on each score report and its format/structure concretely supports the report's purpose and end-user needs. Text/materials developed to support score interpretation (either on reports and/or in ancillary materials) use non-technical language to the extent possible, concrete examples and graphics/illustrations to facilitate understanding and appropriate score use. Limitations related to score interpretation, common misinterpretations, and potential misuses are clearly described. Audience-appropriate reliability/precision information is provided with each reported score (including sub-scores and growth scores) to facilitate the intended interpretations. Devices to support interpretation can include error bars, narrative explanations, numerical examples, graphical representations, interactive displays, categorical determinations, etc. Technical information or concepts such as measurement error or precision are often better served through graphic representations. However, complex graphical representations of data should only be used if there is evidence to suggest that they facilitate understanding. If the assessment measures only a subset of the total universe of standards, it should be made clear which standards are/are not assessed.
5.1: "Test users should be provided with clear explanations of the characteristics, meaning, and intended interpretations of scale scores, as well as their limitation" (p. 102).	
6.10: "When test score information is released, those responsible for testing programs should provide interpretations appropriate to the audience. The interpretations should describe in simple language what the test convers, what scores represent, the precision/reliability of the scores, and how scores are intended to be used" (p.119).	

The complete list of the standards referenced for each of the criteria is included in Appendix B.

An examination of the table in Appendix B reveals that of the thirteen categories of standards,

nine were used to inform the development of the Criteria Evaluation Framework. The categories of standards used to support this effort are shown in Table 2.

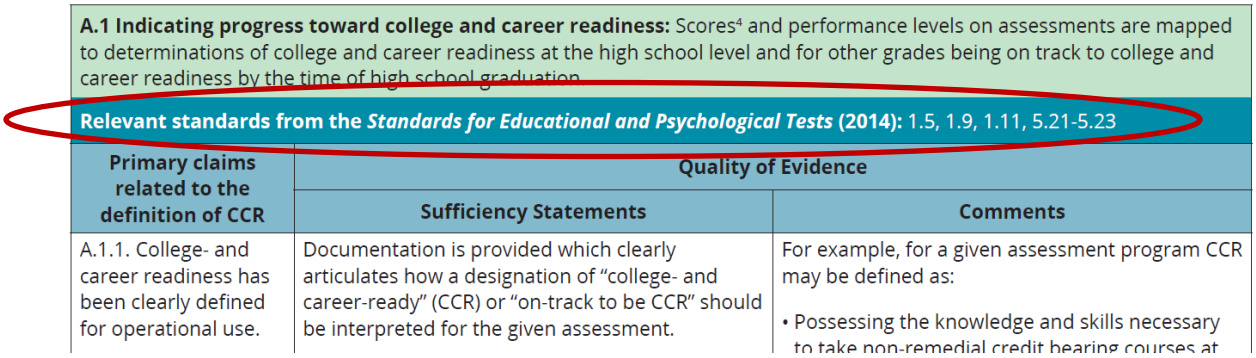
Table 2
Categories of Standards used to inform Criteria Evaluation Framework

Category	Referenced in Criteria Evaluation Framework
Validity	✓
Reliability/Precision and Errors of Measurement	✓
Fairness in Testing	✓
Test Design and Development	✓
Scores, Scales, Norms, Score Linking, and Cut Scores	✓
Test Administration, Scoring, Reporting, and Interpretation	✓
Supporting Documentation for Tests	✓
The Rights and Responsibilities of Test Takers	✗
The Rights and Responsibilities of Test Users	✗
Psychological Testing and Assessment	✗
Workplace Testing and Credentialing	✗
Educational Testing and Assessment	✓
Uses of Tests for Program Evaluation, Policy Studies, and Accountability	✓

In addition to being used to develop help the language of the Criteria Evaluation Framework, we used the *Standards* repeatedly during the extensive review processes that took place between October 2014 to January 2016 to ensure that the standard of quality, or the "bar" set by the sufficiency statements, was aligned with the expectations of the professional *Standards* which govern this work. Our intention was to ensure that the expectations defined by the Criteria

Evaluation Framework were not more stringent or lenient than the standards of practice established by the joint *Standards*. The alignment between these two documents is intended to contribute to a common understanding of what constitutes high quality within the field.

Including the *Standards* within the development and review of the Criteria Evaluation Framework established both the credibility and defensibility of this document and the overall methodology. When stakeholders suggested an expectation was unrealistic or a bar was set too high, not only did we have our own rationale for the requirement, but we could typically point to the *Standards* for external validation of our definition of high quality. Furthermore, we recommend the *Standards* be used as a companion to the Criteria Evaluation Framework when additional detail regarding best practice is desired. For this reason, those *Standards* referenced in the development of the Criteria Evaluation Framework—or are in some other way related to the content of the criterion—are listed explicitly below each criterion (see Figure 4 for an example).



A.1 Indicating progress toward college and career readiness: Scores[4] 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.

Relevant standards from the *Standards for Educational and Psychological Tests* (2014): 1.5, 1.9, 1.11, 5.21-5.23

Primary claims related to the definition of CCR	Quality of Evidence	
	Sufficiency Statements	**Comments**
A.1.1. College- and career readiness has been clearly defined for operational use.	Documentation is provided which clearly articulates how a designation of "college- and career-ready" (CCR) or "on-track to be CCR" should be interpreted for the given assessment.	For example, for a given assessment program CCR may be defined as: • Possessing the knowledge and skills necessary to take non-remedial credit bearing courses at

Figure 4. Example of how relevant standards are listed within the Criteria Evaluation Framework

Going beyond the *Standards*

Just as all of the categories of *Standards* were not utilized in developing the Criteria Evaluation Framework, in some areas the *Standards* did not provide as much guidance as we felt was necessary for evaluation of state operational practices. Specifically, the criteria that

necessitated additional sources beyond the *Standards* were A.5, relating to accessibility for students with disabilities and English learners, A.7, relating to the protections of personally identifiable information, and E.1, relating to test security.[6] For Criterion A.5, the National Center of Educational Outcomes (NCEO) 2007 document entitled *Hints and Tips for Addressing Accommodations Issues for Peer Review*, provided ideas of the types of evidence states should be collecting to support the use of accommodations for their assessment programs. In addition to specific sources of evidence, the document also lists best practices in the use of accommodations to provide for valid inferences about what students know and can do. While we relied heavily on the fairness chapter in the *Standards* to develop the language for Criterion A.5, the level of detail provided in this NCEO document was a useful supplement. Similarly, while standards 6.14-6.16 provide guidance on expectations for maintaining confidentiality of individually identifiable data, operational practices for achieving the level of protection described in the *Standards* are not provided. To support the development of Criterion A.7, relating to the protection of personally identifiable information, we relied on a Statewide Longitudinal Data Systems (SLDS) Technical Brief put out by the National Center for Education Statistics (NCES) in 2010 called, *Data Stewardship: Managing Personally Identifiable Information in Electronic Student Education Records*. Lastly, another document published by NCES as a result of a 2013 Testing Integrity Symposium provided guidance on operational best practices regarding the maintenance of test security that was useful for the development of the Criteria Evaluation Framework for Criterion E.1. This document provides detailed recommendations for the prevention, detection, and response to testing irregularities, with special attention to computer-based tests.

[6] This paper was reviewed by one of the lead author of the *Standards*, Barbara Plake. Dr. Plake argued that there are missing relevant standards from chapters 8 and 9 that could have been used to support Criteria A.7 and E.1. In future versions of our evaluation materials we hope to add these standards to the Criteria Evaluation Framework.

While the *Standards* was a central resource in developing the language of the Criteria Evaluation Framework associated with each of the CCSSO Criteria, in some cases it was necessary to look outside this document for requirements aligned to the purpose and goals of our methodology. As a result, additional reference documents related to accommodations, the protection of confidentiality, and test security, informed the development of the Criteria Evaluation Framework. We believe the evaluation methodology and associated framework are stronger and more useful for evaluating the quality of evidence submitted in these areas given the information provided by these resources.

Added value of the Criteria Evaluation Framework to the use of the *Standards*

Just as we believe the *Standards* add value to our materials, we too hope that our evaluation methodology and framework illustrate a useful way to operationalize the *Standards* for a given purpose or use. We believe this operationalization adds value to the *Standards* in four primary ways, 1) it adds necessary detail to support consistent interpretations for a particular evaluation use case, 2) it ties standards together in a coherent way to support a comprehensive evaluation (again, specific to the particular use case), 3) it explicitly brings the *Standards* to the forefront of an evaluation process in order to increase the reach of the document to people who may not be previously familiar with the standards (e.g., people within state departments of education), and 4) it defines a structure (see Figure 2) that may serve to be a model for other use cases of the *Standards*. Each of these benefits is briefly described in what follows.

While the *Standards for Educational and Psychological Testing* (2014) was not the primary document on which the evaluation tools were built, instead we relied on CCSSO's *Criteria for Procuring and Evaluating High Quality Assessments*, we do feel that the evaluation methodology and Criteria Evaluation Framework add value to the *Standards* by illustrating how

to use and interpret the *Standards* in the context of an comprehensive assessment evaluation. The Criterion Evaluation Framework adds a level of specificity to the *Standards* and enough supporting detail to support consistent judgments about the quality of evidence provided. While *Standards* may certainly be interpreted differently depending on the context in which they are being used, the Criteria Evaluation Framework provides a clear interpretation for one particular use: evaluating state assessments measuring college- and career-ready standards. Secondly, the format of the framework and the intended review procedures provide a coherent structure for evaluating assessments in a systematic, yet holistic, way. The Criteria Evaluation Framework pulls the *Standards* together in a way that supports application for a particular use case. It is worth noting here that the CCSSO Criteria are not universal but rather specific for the particular intended use. Had the purpose of the evaluation been different, the criteria and associated standards would also necessarily be different.

Additionally, the inclusion of the *Standards* in the body of the Criteria Evaluation Framework highlights the central importance of this document for any comprehensive assessment evaluation. While explicitly referencing the *Standards* certainly adds value to the evaluation process as described in the previous section, we also hope that the Criteria Evaluation Framework can help spread the reach of the *Standards* to those who might not otherwise be familiar with the document. The final section of this paper, entitled "Benefits of the Claim-Based Structure" provides more detail about the added value of this structure for the interpretation, use, and coherence of our professional *Standards*. Because this fourth benefit stands alone, in that it applies beyond the context of the particular use case for which the test characteristics evaluation methodology and Criteria Evaluation Framework are intended to serve, the benefit of the claim structure is described in detail in its own section.

Benefits of the Claim-Based Structure

Prior to developing the test characteristics methodology the Center team took time to articulate the goals for the methodology—how we intended for information resulting from the methodology to be used, including who the consumer of those results might be—and any constraints or guardrails that could impact the design of the process and tools (e.g., resources, time, required expertise). These elements served as the foundation underlying the design of the methodology and solidified our need for a tool by which to operationalize the criteria for evaluation (i.e., the Criteria Evaluation Framework). The organizing structure of the Criteria Evaluation Framework, in terms of claims, sufficiency statements and comments, was influenced both by these goals/assumptions and a few additional factors which emerged early on in our discussions about the methodology, including:

1. **The breadth of the CCSSO criteria** – the requirements underlying each criterion were numerous and left significant room for interpretation. To establish a coherent set of parameters for evaluation we knew it would be beneficial to partition the criteria into manageable elements or parts.

2. **Our desire to structure the evaluation process in terms of an interpretive argument** – where the evidence required to inform the evaluation of a given assessment stemmed directly from the inferences and assumptions necessary to support the quality of that assessment.

3. **The impact of contextual factors (e.g., the assessment's phase of development, mode of administration, intended use of results, etc.) on decisions regarding the relevance and quality of evidence provided to support evaluation**. Since such factors could never be fully accounted for in an evaluation tool, we knew that comments and examples would be necessary to highlight those areas where contextual factors might impact how evidence was evaluated.

The decision to represent each criterion in terms of claims was a direct result of the first two points identified above. This format proved beneficial for a variety of reasons. First, the claim-based structure required us to be specific about the inferences underlying each of the criterion

statements—which can be thought of as statements of goals and values for the assessments to be evaluated. This is important because different people can have different conceptualizations of what assessment quality means, both in general and relative to the specific areas represented by the CCSSO criteria (e.g., validity, reliability, security, etc.). Being explicit and transparent about the claims underlying each criterion was necessary to support consistency in the interpretation and evaluation of each criterion within the context of an assessment evaluation.

Second, the claims provide both a foundation for identifying necessary evidence and a rationale for why a given piece of evidence should be considered. The sets of claims were developed to articulate those conditions which must hold to lend support to a given CCSSO quality criterion given the overarching goal of evaluating the quality of a summative assessment aligned to college and career-ready content standards. The claims represent expectations specific to this use-case which allowed us to be more specific and purposeful in defining the type of evidence expected. This was especially the case in thinking about evidence necessary to support the claims associated with criteria A.1, A.2, and D.1 which reflect the intent of the assessments as providing for inferences related to college- and career-readiness. For example, instead of just asking for evidence regarding the procedures used to establish the performance standards, we asked for evidence that the procedures, materials and data used to establish the performance standards would support score-based inferences regarding the readiness of students for college and careers (however readiness had been defined). Of course some claims were more general and would generalize across a variety of use cases. For example the claims associated with A.7 would likely generalize to any context in which security of assessment results was a concern.

The more specific the details of the use case, the more intentional one can be in articulating claims and the more specific one can be about the evidence necessary to support

those claims. The Criteria Evaluation Framework represents an application of this claim-based structure to one broad use case—the evaluation of a standardized, summative assessment measuring college- and career-ready standards. The tool is intended to be used by the broadest audience possible, which influenced the assumptions made and, consequently, the level of detail afforded. If we had been developing the framework to evaluate a particular assessment in a known state context, for which the purpose and intended use of the assessment was clear and well documented, we could have been even more specific in the claims, and consequently the desired features of provided evidence.

A model for adding to the utility of the *Standards*.

We argue that the claims-based structure of the Criteria Evaluation Framework adds utility to the *Standards* as it operationalizes the ideas in the *Standards* for a particular use case. We believe that this structure provides a model which can be replicated for other use cases in order to carry out the intention of the *Standards* in a variety of settings. While those in attendance at the NCME annual meeting are likely to be familiar with the *Standards*, knowledge of the guidelines that govern our work does not likely extend far beyond the membership of our professional measurement organizations. For those less familiar with the *Standards*, who are either engaging in an assessment evaluation or consuming the results of an assessment evaluation, claims provide a first step in helping to determine which *Standards* are relevant and who is responsible for providing evidence to support them. Specifying the claims one wants to make within a given context is likely a much more approachable task for non-measurement professional that the task of identifying which *Standards* apply to a particular context and must be addressed. The former should act as a first step in any process as it provides a coherent means to addressing the latter. We believe this process is a way to build "*Standards* literacy," in

way that is inherently useful—to understand how the *Standards* could and should be applied in the consideration of assessment-related evidence for a particular purpose.

In the case of the test characteristics Criteria Evaluation Framework, claims were expressed in terms of the expectations necessary to defend the quality of a summative assessment developed to measure college- and career-ready content standards (as logical extensions of the goals outlined in the CCSSO Criteria). Once the claims were defined, relevant professional standards were identified and used as the foundation by which to identify and articulate evidence of quality in support of those claims as represented in the sufficiency statements. Essentially, the sufficiency statements translated the expectations of the *Standards* for use in a particular context. The examples of evidence listed in the sufficiency statements allowed us to be specific, adding to the usefulness of the necessarily broadly-written *Standards*, while preserving their original intent. This process for articulating the claims and sufficiency statements and the primary sources on which the language of each of those components should be based, is outlined in Figure 5.

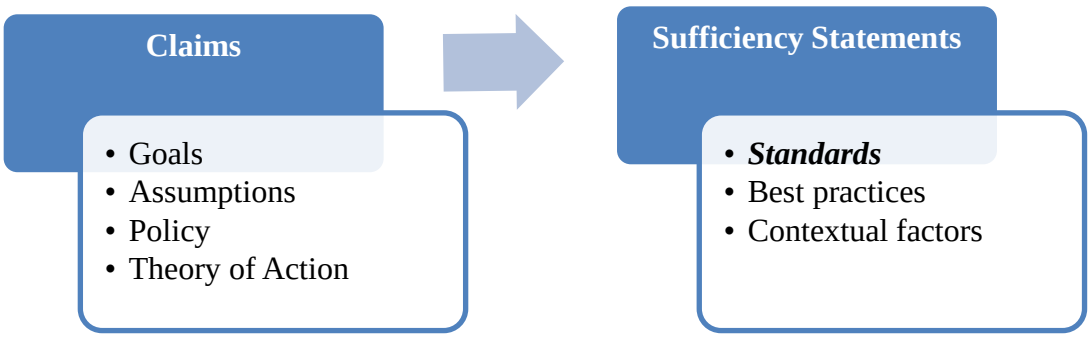


Figure 5. Suggested model for developing a framework based on Standards

Concluding Thoughts

The test characteristics methodology and associated Criteria Evaluation Framework, developed by the Center for Assessment to support evaluation of assessments relative to CCSSO's *Criteria for Procuring and Evaluating High Quality Assessment*, provides one

example of how the *Standards for Educational and Psychological Testing* (2014) have been referenced and used to support the improvement and evaluation of educational assessments. The development of these Criteria Evaluation Framework relied on the *Standards* during both the process of articulating the original language and its many rounds of revision. We believe the *Standards* add great value to our evaluation documents in that they support the defensibility of the expectations outlined in Criteria Evaluation Framework sufficiency statements, and also by providing an additional resource to which we point evaluators when supplemental guidance may be sought. Additionally, we believe our work adds value to the *Standards*; not only for the particular use case for which we developed our materials, but also in the design of the claims-based structure of the Criteria Evaluation Framework. We hope that the Criteria Evaluation Framework can serve as a model for operationalizing the *Standards* that could generalize to be applied to support their many potential other uses.

References

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

Christensen, L.L., Lail, K.E., & Thurlow, M. L. (2007). *Hints and tips for addressing accommodations issues for peer review.* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

CCSSO. See Council of Chief State School Officers.

Council of Chief State School Officers. (2014). *Criteria for procuring and evaluating high-quality assessments*. Retrieved from http://www.ccsso.org/Documents/2014/CCSSO%20Criteria%20for%20High%20Quality%20Assessments%2003242014.pdf

Hall, E. & Lyons, S. (2016). *A guide to evaluating college- and career-ready assessments: Focus on test characteristics – evaluation methodology*. National Center for the Improvement of Educational Assessment (Center for Assessment). Retrieved from http://www.nciea.org/wp-content/uploads/CFA-TestCharacMethod-EvalMethod_Final.pdf

Hall, E. & Lyons, S. (2016). *A guide to evaluating college- and career-ready assessments: Focus on test characteristics – criteria evaluation framework*. National Center for the Improvement of Educational Assessment (Center for Assessment). Retrieved from http://www.nciea.org/wp-content/uploads/CFA-TestCharacMethod-CriteriaEvalFramework_Final.pdf

U.S. Department of Education. (2013). *Testing integrity symposiums: Issues and recommendations for best practice*. Institute of Education Sciences, National Center for Education Statistics. Retrieved from http://nces.ed.gov/pubs2013/2013454.pdf.

U.S. Department of Education. (2010). *Data stewardship: Managing personally identifiable information in electronic student education records*. SLDS Technical Brief #2. Institute of Education Sciences, National Center for Education Statistics. Retrieved from http://nces.ed.gov/pubs2011/2011602.pdf.

Appendix A: Example Claim from the Criteria Evaluation Framework

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 **Relevant Joint Standards (2014):** 6.14, 6.15, 6.16		
Primary claims related to student privacy	**Quality of Evidence**	
	Sufficiency Statements	**Comments**
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.	Procedures and documentation (e.g., training materials) demonstrate that contractors and any subcontractors utilized to support the assessment process, are well trained and compliant with FERPA. All entities that may have access to secure data (at any level) as part of the assessment process are clearly identified. Procedures exist and are documented and actively monitored to comply with FERPA regulations for the security of educational, personally identifying and directory information, as necessary given the type of student data that will be collected and stored in conjunction with the assessment. The assessment vendor (i.e., publisher, developer, provider, or scorer) provides training to employees and monitors/documents compliance to the requirements and prohibitions of FERPA to the extent necessary/appropriate given their specific roles and responsibilities and the data to which they will have access. A process is in place to ensure the sponsoring agency is notified of any security breaches that may result in student data becoming available to non-authorized individuals.	Often, different vendors are acquired for different elements of the assessment process (e.g., development, publishing, administration, scoring, etc.). Contractual arrangements are recommended where vendors are contractually obligated to advise the state in the event of a security breach that involves examinee data, item data or other test-relevant data. It is important to note that many of the requirements related to FERPA will fall under the responsibility of the state, such as providing parents/students with annual information about FERPA and maintaining the privacy of state/school records. *For assessments developed to be used in multiple states, evidence should be provided for any/all procedures utilized to comply with FERPA that are common across all states*.

Appendix B: Standards Referenced within the Criteria Evaluation Framework

CCSSO Criterion	Standards Referenced
A.1 Indicating progress toward college and career readiness: Scores[7] 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.	**Validity:** 1.5, 1.9, 1.11 **Cut Scores:** 5.21, 5.22, 5.23
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.	**Validity:** 1.1, 1.2, 1.6, 1.8, 1.9, 1.11, 1.13, 1.16, 1.17, 1.18, 1.25 **Test Design:** 4.0, 4.1, 4.2 **Design and Development of Educational Assessments:** 12.2,12.4, 12.11 **Design and Development of Testing Programs and Indices for Program Evaluation, Policy Studies, and Accountability Systems:** 13.3
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.	**Reliability/Precision:** 2.0, 2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 2.7, 2.8, 2.10, 2.12, 2.13, 2.14, 2.16 **Design and Development of Educational Assessments:** 12.2
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	**Standard Error of Measurement:** 2.15 **Test Design and Development:** 4.3, 4.4, 4.8, 4.9, 4.10, 4.11, 4.12, 4.18, 4.19, 4.20, 4.21

[7] 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.

(e.g., paper, computer, including multiple computer platforms). • **The score scales** facilitate accurate and meaningful inferences about test performance.	**Interpretations of Scores:** 5.2, 5.6, 5.7, **Score Linking:** 5.12, 5.13, 5.14, 5.15, 5.16 **Educational Testing and Assessment:** 12.3, 12.5, 12.6, 12.8
A.5 Providing accessibility to all students, including English learners and students with disabilities. •**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. •**Offering appropriate accommodations and modifications:** Allowable accommodations and modifications[8] 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. •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.**	**Fairness:** 3.6, 3.7, 3.8, 3.9, 3.10, 3.11, 3.12, 3.13, 3.14,3.15, 3.17 **Supplemental Resource:** Christensen, L.L., Lail, K.E., & Thurlow, M. L. (2007). *Hints and tips for addressing accommodations issues for peer review.* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
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	**Reporting and Interpretation:** 6.14, 6.15, 6.16 **Supplemental Resource:** U.S. Department of Education. (2010). *Data stewardship: Managing personally identifiable information in electronic student education records*. SLDS Technical Brief #2. Institute of Education Sciences, National Center for Education Statistics. Retrieved from http://nces.ed.gov/pubs2011/2011602.pdf.

[8] The 2014 Standards for Educational and Psychological Testing define modifications as changes that impact the construct, whereas accommodations preserve the construct. Henceforth, the Criteria Evaluation Framework upholds this definition of these terms.

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.	**Validity:** 1.3, 1.14, 1.15 **Interpretations of Scores:** 5.1, 5.2 **Reporting and Interpretation:** 6.10 **Educational Testing and Assessment**: 12.11, 12.18
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	**Reporting and Interpretation:** 6.13 **Administration, Scoring, and Reporting of Educational Assessments:** 12.19
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.	**Design and Development:** 4.5, 4.15, 4.16 **Test Administration:** 6.1, 6.2, 6.4, 6.5, 6.6, 6.7 **Test Security and Protection of Copyrights:** 9.21, 9.22 **Educational Testing and Assessment:** 12.7, 12.16 **Supplemental Resource:** U.S. Department of Education. (2013). *Testing integrity symposiums: Issues and recommendations for best practice*. Institute of Education Sciences, National Center for Education Statistics. Retrieved from http://nces.ed.gov/pubs2013/2013454.pdf.