A GUIDE TO EVALUATING COLLEGE- AND CAREER-READY ASSESSMENTS:

Focus on Test Characteristics

Criteria Evaluation Framework

MARCH 2016

ERIKA HALL, PH.D.
Center for Assessment

SUSAN LYONS, PH.D.
Center for Assessment

Center for Assessment

THE CRITERIA EVALUATION FRAMEWORK

The Criteria Evaluation Framework (CEF) is a tool developed to support the evaluation of assessments against CCSSO's Criteria for Procuring and Evaluating High Quality Assessments[1]. The CEF was designed to support the evaluation of those criteria associated with test characteristics; that is, those reflecting the psychometric and statistical properties of assessment instruments and the quality of test administration, reporting, and any supplemental information provided to aid in the interpretation and use of test results. For each criterion, the framework lists several claims which should be satisfied and provides examples of high quality evidence that would lend support to those claims. The complete test characteristics evaluation methodology defines each component of the framework and describes the manner in which the tool may be used to support a comprehensive assessment evaluation[2]. While the Criteria Evaluation Framework may be referenced independent from the evaluation methodology, the reader is strongly encouraged to review the CEF overview provided in the test characteristics methodology to fully understand the intent and structure of CEF and its elements.

[1] 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

[2] See the companion document to this framework entitled "A Guide to Evaluating College- and Career-Ready Assessments: Focus on Test Characteristics – Evaluation Methodology."

CCSSO Criterion A.1

The evaluation of evidence associated with A.1 involves judging the degree to which the documentation provides evidence that the assessment scores support determinations of college and career readiness or being on-track to college and career readiness. The primary claims related to this criterion are divided into three main sections:

1) Readiness definition: claim A.1.1 evaluates whether, and how clearly, college and career readiness has been defined for the given assessment program.
2) Performance Level Descriptors: claims A.1.2-A.1.5 evaluate the quality of evidence related to the performance level descriptor development process.
3) Standard setting process: claims A.1.6-A.1.11 evaluate the quality of evidence related to the standard setting process[3] and results.

Additionally, because the integrity of the performance standards (a.k.a. cut scores) depends on the reliability and accuracy of scaling and equating procedures, a secondary set of claims from criterion A.4 are appended to support a holistic judgment regarding criterion A.1.

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. Any limitations or restrictions associated with a given definition of college- and career-ready are articulated.	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 the start of college. • Performing at a level of proficiency (in the content area) that represents a high probability of earning a C or better in related first year college courses • Displaying those knowledge and skills representing CCR as defined by the expectations (Performance Level Descriptors) associated with this standard. Similarly, on-track to be CCR may be defined as: • Performing at a level consistent with that necessary to meet the CCR benchmark in high school if maintained.

[3] The standard setting *process* includes the standard setting meeting as well as any planned processes/judgments which lead up to the final approval. In contrast the standard setting *methodology* refers to the specific technique or approach used by panelists to recommended performance standards within the context of the standard setting meeting.

[4] 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.

		• Displaying those knowledge and skills representing on-track for CCR as defined by the expectations (Performance Level Descriptors) associated with this standard Because pre-existing definitions of college and career ready will vary by institution, the assessment program must adopt an explicit definition of college- and career-ready which includes enough detail to provide for common interpretations of performance relative to this standard for all who use a given assessment.
Primary claims related to the performance level descriptors[5]	**Quality of Evidence**	
	Sufficiency Statements	**Comments**
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.	The PLD development and articulation process uses CCR content standards as the basis for developing coherent expectations associated with student performance at each performance level within and across grades (e.g., vertical articulation). The PLDs are built directly from the CCR content standards in that the expectations are defined relative to both content knowledge and cognitive processes. The process focuses not only on the coherence within a particular grade/content area, but also on the consistency of expectations across grades levels, especially, with respect to defining progress towards college and career readiness. Materials and instructions for developing PLDs consistently focus educators back to the assessed content standards, the definition of college and career readiness and the manner in which they will be jointly addressed on the assessment (e.g., multiple choice, constructed response) as reflected in test blueprint and/or item specification documentation. Documentation is provided which clearly illustrates the process for ensuring accurate and adequate alignment between the PLDs and the CCR content standards (or domains/clusters when applicable) within and across grades. Materials and instructions articulate how the level of	The PLD development process should not allow for the development of expectations associated with content standards that are not targeted for inclusion on the assessment (e.g., speaking and listening). Likewise, if test assembly is done at the domain level rather than the standards level, so should the development of the PLDs. Panelists should always be referred back to the test blueprint and any specifications that detail the content limits associated with standards to be assessed. Similarly, expectations should be written in consideration of the way in which a student will be asked to demonstrate a particular skill/competency within the context of the assessment. If CCR and/or on-track performance standards are established using completely empirical procedures (see Footnote 10) and these standards are also intended to support criterion-referenced interpretations, documentation should indicate the procedures used to establish PLDs that align to the standards and represent the content expectations defined by these standards.

[5] Note: Claims A.1-A.5 are based upon the assumption that Performance Level Descriptors will be generated and subsequently used to support standard setting. However, if CCR is defined in terms of an external validity criterion (e.g., likelihood of success in credit-bearing courses) and a reliable and valid external criterion exists by which to estimate cut scores in light of this definition, PLDs and standard setting may not be required to establish the CCR performance standard (a.k.a., cut score). Similarly, if on-track is defined in terms of a point on the scale that serves to predict attainment of the on-track or CCR performance standard at the subsequent grade, PLDs may not be used to establish on-track standards. In these situations PLDs may be generated after the fact, using the standards as an anchor by which to establish content-based descriptors of CCR, OR a separate standard setting may be conducted to establish performance standards that represent criterion-based descriptions of differentiated levels of performance. In the latter case, the CCR or on-track standard would be a stand-alone, empirically derived standard which is not used to represent the transition between performance levels.

	proficiency represented in the CCR content standards is intended to map to different performance levels. If content standards are written to represent what a "Proficient" or "CCR" student should be able to do, for example, this should be stated in advance and be clearly represented in the PLD development process.	
A.1.3. Knowledgeable experts were involved in the process of developing and reviewing the PLDs.	Representatives from grade-level educators, higher-education, career and technical education, and industry (e.g., local employers hiring high school graduates) are involved in the specification and/or review of performance level descriptors. Representatives include those affiliated with different types of institutions e.g., 2 and 4 year universities, career and technical schools, and those having appropriate content expertise in the subject area. Grade-level educators involved in developing PLDs include representation from those who work with all types of students (e.g., English learners and students with disabilities), and/or who were in these groups when they were students. Documentation is provided indicating the direct involvement of one or more technical experts in the review and approval of the PLD development methodology and results. Provided materials discuss not only who was involved, but indicate what was reviewed, the manner/type of feedback received.	The number of representatives of CTE, industry and higher education that are appropriate (and they role they play) may vary depending on the grade and content area within which PLDs are being established. For example, HS assessments used specifically to make final determinations of CCR should reflect greater representation of HE. Similarly, mathematics assessments that address foundational skills necessary for success in a broad range of technical fields should include industry or CTE representation in either the development or review of PLDs. The quality of the evidence presented will depend on who was involved in the review of the PLD process. External reviewers are preferable to internal reviewers. Examples of highly qualified external experts would be those in the field of educational measurement who have demonstrated substantial experience running PLD development meetings and/or have a record of publications in peer-reviewed journals about setting standards or related measurement topics. However, internal, independent reviewers are preferable over less transparent quality control procedures.
A.1.4. The process used for developing performance level descriptors (PLDs) supports their intended use(s).	The PLD process clearly identifies all of the ways in which the PLDs are intended to be used (e.g., support inferences regarding student achievement and progress, inform standard setting, support future item development, support instruction by clearly defining expectations for student performance, etc.). If PLDs are intended to serve multiple purposes the process reflects a clear connection between PLD development, each purpose and the assessed content standards. Documentation is provided that indicates that panelists were informed of how the PLDs were to be used as part of training.	For example: PLDs developed to support item development should be extremely detailed and written at a fine grain-level. The process should allow for the design of items that not only target distinctions between PLDs but also span the score scale so as to better determine "on track." PLDs written to support reporting should be broad, yet useful and informative for the intended audiences about what students know and can do. PLDs written to provide educators with an indication of the type/level of skills represented by students at different performance levels may be detailed, but not all inclusive.

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.	Evidence is provided to show that skills and skill levels described in the PLDs are aligned with the KSAs assessed by the items that most highly discriminate performance at each respective achievement level. The PLD development process includes a plan for re-evaluation of the PLDs as needed to account for factors that may invalidate the original statements, such as: a change in the range or type of content and cognitive processes to be assessed (i.e., change to test blueprint), the use of a standard setting process that prioritizes empirical data over content-based judgments, or new evidence of skills and content knowledge important to college- and career-readiness.	If, for example, PLDs are intended to be criterion referenced, and are written before standard setting, the accuracy of the expectations for student performance should be validated after performance standards are put in place – which is especially the case when external/impact data are the primary means by which performance standards are established. If PLDs are developed in light of empirically derived CCR or on track performance standards (i.e., cut-scores), the process used to map these cut-scores back to the content standards through test items provides evidence in support of this claim. For example, re-evaluation may include reviewing the skills associated with a sample of items that are representative of the performance level range (i.e. the item parameters are within the surrounding cut scores) to ensure they align with the expectations defined in the performance level descriptors.
Primary claims related to standard setting	**Quality of Evidence**	
	Sufficiency Statements	**Comments**
A.1.6. A description and coherent rationale are provided for how the proposed and/or implemented standard setting methodology[6] yields valid determinations of progress toward, or attainment of, college and career readiness.	The rationale for the standard setting methodology is clearly provided. The standard setting process and materials are appropriate given the definition of college-and-career readiness (as reviewed in A.1.1.A) and the inferences scores are intended to support. The methodology provides for performance standards (a.k.a. cut scores) that are coherent across grade levels. If multiple inferences are intended (e.g., predictive, growth, and criterion-referenced), the methodology describes how recommendations suggested by disparate inferences are prioritized, weighted and resolved, and outlines the rationale for those procedures and decisions. If standard setting panels are convened for only for a few grades (e.g. 4, 8 and 11), a sound, appropriately vetted rationale is provided for selecting those grades as well as the procedures/ techniques used to interpolate/extrapolate recommended performance standards (a.k.a. cut scores) to the tested grades not represented by panels (as necessary).	A variety of standard setting methodologies exist in the literature (e.g., Bookmark, Briefing Book, Angoff, Contrasting Groups, etc.). There are pros/ cons associated with each methodology and the appropriateness of each may be influenced by a variety of factors, including: the context in which the standard setting occurs, its purpose, definition of the standard, who is involved and a variety of other factors. The impetus behind the specific methodology selected or developed for use in light of the specified definition of the standard should be transparent and clearly articulated. The standard setting methodology and data necessary/appropriate to inform it will vary depending on the manner in which "readiness" or "on track to be ready" are defined. For example, readiness may be intended to reflect: - The likelihood of obtaining a given score, or performing at a particular level on a criterion measure - The point on the reportable scale where a student has "just enough" knowledge and skills to be CCR, as defined by the PLDs, etc."

[6] 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.

	If the methodology is newly developed or represents a departure from best practice, the rationale for any modifications made is provided. Documentation is provided which shows that a panel of technical experts was involved in the review and approval of the standard setting methodology and proposed implementation; Such documentation details not only who was involved, but what was reviewed, the manner/ type of feedback received, and any actions taken based on that feedback.	Similarly "on track to be ready" may be defined in terms of: - The likelihood of meeting on-track standards in the next grade - The point on the reportable scale where a student has "just enough" knowledge and skills to be on-track, as articulated by the PLDs, etc." In the first instance empirical evidence that illustrates the relationship between different points on the scale and the criterion measure will take precedence; in the latter definitions, judgment-based standard setting procedures based on test content (e.g., Bookmark; Angoff, etc.) are more likely to be appropriate, but should also be tied to statistical projections of 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.	Representative individuals from grade-level educators, higher-education institutions, industry, and career and technical education are involved in the recommendation and/or evaluation of performance standards. Representatives include those affiliated with different types of institutions e.g., 2 and 4 year universities, trade schools, and those having appropriate content expertise in the subject area. The intended contribution (content expertise, representativeness, special interests, etc....) of each standard setting participant to the overall process is clearly articulated. The process used to identify panelists for inclusion in the standard setting process is clearly described. Grade-level educators involved in standard setting include those in the best position to represent special groups of interest (e.g., English learners and students with disabilities).	For example: representatives who serve as higher education administrators may not be appropriate to include in a test-driven standard setting process, but could be involved in a review/ evaluation of the expected impact associated with proposed cut. For example, to identify standard setting panelists, specific districts may be targeted initially to ensure "representativeness." Those selected to represent special interest groups may include teachers who work with this student population, teachers who were in these groups when they were students, or others defined by school administrators as most qualified to fulfill this role.
A.1.8. Appropriate external CCR benchmarks and research studies are/ were used in the standard setting process.	The rationale underlying the range and type of external benchmark data and research studies presented to support standard setting is clearly articulated and includes the factors/evidence considered when making decisions regarding which evidence to include/exclude. If external benchmarks are not part of the standard setting process a rationale is provided to support this decision. The standard setting process includes a description of how and when external benchmarks and studies are introduced into the standard setting process.	Examples of sources of external evidence include: student performance on current state assessments, NAEP, TIMSS, PISA, ASVAB, ACT, SAT, results from state assessments such as Smarter Balanced and PARCC, relevant data on post-secondary performance, remediation, and workforce readiness. Factors considered in the selection of external benchmarks may include: the content alignment of benchmark measures to the assessment; the inferences the benchmark measure was intended to support (e.g., CCR), the population of students to which the benchmark is administered (e.g., international sample, graduating HS seniors,

	The manner in which panelists are intended to use and prioritize external evidence (alone and in conjunction with PLDS) when making recommendations is clearly articulated, consistent with intent of the performance standards, and reasonable given the technical quality and relevance of the measures. The quantity of external benchmarks does not necessarily equate to quality, the cognitive load on the panelists needs to be considered. Feedback provided by standard setting panelists suggests that they understood the data (e.g., impact data) provided and the manner in which it was intended to be used to support the standard setting process. Panelists indicate that they are satisfied with their performance standard (a.k.a. cut score) recommendations at the end of the standard setting process.	etc.…), the psychometric properties of the benchmark measure, and the clarity and understandability of evidence for panelists or others considering it. Which factors are important will depend on the role the benchmark is intended to serve. External data that are not relevant, of poor quality, or repetitive can do more damage than good – therefore sufficiency of evidence should not be based on the amount of external evidence provided, but importance and usefulness of that evidence. External benchmarks should not be provided within the context of the standard setting process unless the value they add and the role they are intended to play is made explicit. When provided with multiple pieces of evidence, panelists must be given information and instruction that helps them to weigh and prioritize each piece of evidence in making cut-score recommendations, especially if they vary in terms of quality and relevance. Whether the weighting is prescribed, or whether information is given to help panelists make their own weighting decisions should be made clear.
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.	Procedures/techniques used to smooth the final set of recommended CCR cut scores for all grades/within a content area are clearly defined and accompanied by a coherent rationale. If smoothing process moves the proposed cut scores, a process is in place to ensure/validate that the movement did not alter the intended meaning of the standard. If cut scores are moved after the standard setting meeting, a reasonable rationale is provided for these changes that align with the definition of CCR. Changes or proposed modifications to cut scores which occur after one or more years after implementation are supported by student performance or validation studies which show such movement is necessary to support intended uses/interpretations.	There are a variety of adjustments that may be made to recommended cut scores throughout the standard setting process. For example: cross-content smoothing, cross-grade smoothing, policy adjustments after panelists leave, and post-stabilization evaluation and/or adjustment of cut scores. Descriptions should explain how decisions will be made at each step and include a rationale for any decisions made, if applicable.

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.	As part of a comprehensive validity evaluation[7], a set of short-term and longitudinal studies is proposed to evaluate the validity of the inferences the performance standards are intended to support. Studies are directed at evaluating the validity of college-and-career readiness (or on-track to CCR) inferences and the information collected goes beyond assessment data to other indicators of CCR. Studies defined to support the validity of CCR performance standards and associated inferences include the collection and review of high quality empirical data that is consistent with the manner in which "readiness" has been defined and is consistent across grades. The sampling plans for validity studies are included with the description of the study methodologies along with an accompanying rationale for the plan.	If, for example, readiness is operationalized in terms of a given probability (e.g., 67%) of attaining a grade point of B- or greater in a related, non-remedial college credit bearing course, validity studies should focus on collecting evidence that supports the validity of the standard for this purpose. If the standard is operationalized in terms of "expected knowledge and skills" as defined by the PLDs, then data should be collected that shows that the performance of students falling within a given performance level is consistent with those expectations. For example, this consistency could be reflected in expected or observed impact data, or estimated probabilities for success on criterion measures. *For assessments developed to be used in multiple states (e.g., consortia-based test, ACT, and SAT), the studies (including sampling plans and rationales) should address and account for the different policy and population contexts of each of the states administering the tests.*
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.	Documentation is provided that indicates the standard setting was conducted as intended. Any deviations in the planned procedures are accompanied with rationales and evidence that the validity of the performance standards was not sacrificed. Evidence is provided indicating how/and to what extent the resulting performance standards and the results of validity studies were reviewed by qualified technical experts. When procedures or analyses are required to translate panelist ratings to the reportable scale metric, evidence is provided to verify that that these calculations were performed correctly.	For example, documentation that the standard setting was conducted as intended may include: • A report from an external evaluator at the standard setting meeting which indicates the plan was implemented as proposed. • A detailed summary of the standard setting process as reported by a qualified independent observer (or panel) followed by a summary of how that process adhered to (or differed from) the proposed plan. • An independent analysis or QC report which verifies that cut-scores recommendations were calculated accurately and in the manner intended. Evidence that the performance standards and associated validity studies were reviewed may include: • Agendas, meeting minutes and materials developed and presented to support evaluation and review. • A list of those involved in the expert review panel. • A summary of feedback, recommendations or approvals obtained in light of expert review. • A summary of any actions taken based on expert feedback and recommendations. The quality of the evidence presented will depend on the extent to which the experts are independent from the standard setting process,

[7] CCSSO Criterion A.2 evaluates the quality of the evidence provided related to the comprehensive validity evaluation plan.

		and the amount/rigor of evidence reviewed. External reviewers are preferable to internal reviewers. However, internal, independent reviewers are preferable to less transparent quality control procedures.
Secondary claims from A.4 related to scaling and equating	***Quality of Evidence***	
	Sufficiency Statements	***Comments***
A.4.6 – A.4.10	*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 A1.6 and A1.10). 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.4.8-A.4.13, 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.*	

CCSSO Criterion A.3[8]

The evaluation of evidence associated with criterion A.3 involves judging the degree to which the provided documentation can support the quality of the reliability analyses and results to support the intended uses and interpretation of scores. In addition to the primary claims relating to reliability procedures and results (A.3.1-A.3.3), a secondary claim from criterion D.2 has been appended due to the evidence associated with this claim relating to informing users of the magnitude of error surrounding each reported score. The primary and secondary claims should be considered together when making a holistic judgment regard criterion A.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.

Relevant Joint Standards (2014): 12.2, 2.0-2.8, 2.10, 2.12-2.14, 2.16

Primary claims related to reliability	Quality of Evidence	
	Sufficiency Statements	**Comments**
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.	For all reported scores (e.g., total scores, subscores, cut scores, growth scores, predicted scores) reliability coefficients are calculated for the overall student population and for each reported sub-population (e.g., overall, race/ethnicity, gender, English language proficiency, disability status, economic disadvantage status, and grade level, performance level). Rationale for the reliability indices selected for use are included for each score: • Depending upon the psychometric model that is being used (classical or IRT or another model), the type of reliability index is justified. • An appropriate type of reliability index is reported for each type of score (total, subscore, process, classification, growth) that is being reported. This information is reported for all subgroups of interest. • The reliability of assessments that use multiple item formats reflects the format variation and how different types of items contribute to or detract from the overall test reliability. If scores on performance tasks, e.g., writing, contribute to overall scores then how scores from those tasks are folded into an overall reliability is described. The descriptions for quantifying/calculating reliability indices are clear, expressed in terms of statistics appropriate to each method. • The descriptions include the type of reliability indices and standard errors to be calculated, the formula or methodology used, and any adjustments made for restriction of range or variability. If census data are not used, a description and rationale related to the sampling procedure and generalizability of sample is provided. • When significant variations are permitted in test administration procedures, separate reliability analyses are provided for scores produced under each major variation.	

[8] Please note that because Criterion A.2 is to be evaluated only after all other criteria have been considered, the CEF language developed to support Criterion A.2 is located at the end of this document.

	• When human judgment enters into scoring, procedures and methods for gathering and evaluating inter-rater, and within-examinee score reliability are provided. The impact of factors such as lesser precision on the subjectively scored items is documented with its impact on overall and subscore reliability estimates. If constructed-response items are scored locally, than reliability indices specific to these items are calculated at the local level and evaluated holistically at the program level. • If scores are reported in a manner that invites comparisons across scores or subscores, (e.g., differential performance across claims, sub-claims or targets) then methods for evaluating the precision of the difference scores are presented and implemented.	
A.3.2. Clear criteria are in place for evaluating the appropriateness of obtained reliability indices and estimates of precision.	Rationales for the specified criteria for uncertainty are in alignment with each of the intended interpretations, uses and potential sources of error for a given score (e.g., reliabilities for norm- and criterion-referenced interpretations of a score are given separate consideration). Score reliabilities that are extremely low may signal scores that are inappropriate for their intended use. Rationales include specific attention to the tension between prioritizing precision at the CCR and on-track to CCR cut scores and also providing reliable scores for essentially all students. Test information is most critical at the cut scores while also maintaining accurate assessment for students at the extremes of the assessment scale. Criteria for evaluating the adequacy of reliability and precision indices account for the numbers of items/tasks/pieces of evidence necessary to support reporting and intended inferences. Acceptable levels of reliability for items that require subjective scoring are articulated and required prior to the inclusion of such items into a reported score. Factors that may influence the attainment of these criteria (e.g., first year of the administration, use of field-test data, distinctness of sub-scores, administration conditions such as speededness, minimal training sets, the degree to which educators are prepared in the content of implementation or student exposure to effective instruction, and the rapidity of curriculum implementation) and their implications are clearly documented along with plans for addressing such deficiencies. Procedures are in place to investigate the cause and potential implications of reliability estimates that fall outside the desired or expected range. Plans are specified to improve the reliability of scores where needed, in a timely manner.	Minimal values for reliability are context dependent, but a general rule of thumb is that the minimum score reliability for low-stakes use is generally around .80, and around .90 for higher stakes. Likewise, acceptable magnitudes of standard errors of measurement (conditional and overall) will change with the intended interpretations and uses. Scores that are used to make high-stake decisions will necessitate smaller standard errors. Similarly, the standard errors of measurement should likely be lowest near the cut scores. Score precision can be expressed as the frequency of classification errors, with statements like the following: At least 80 percent of the students classified as proficient are expected to have true scores in the proficient range. In many cases, estimates of the frequency (and consequences) of classification errors will be more meaningful to policy-makers than simple confidence bounds.

A.3.3. The pre-specified reliability and precision indices were estimated and the results indicate adequate support for intended uses.	Documentation is provided showing that planned reliability and precision indices were calculated and that the results adequately support the intended uses for essentially all students. Provided documentation may include: • Representative samples of observed or estimated reliability indices and precision coefficients for total scores and sub-scores, classification consistencies, and precision at the cut scores. • Results from generalizability analyses conducted to evaluate the contribution of different factors (e.g., subgroups, schools, test forms, raters) to the error of test scores/sub-scores. • Evidence that the reliability and precision estimates meet the criteria specified for adequacy (see A.3.2) and/or that estimates that do not meet the criteria are reasonable and that plans to improve those estimates are in place or that there are sound policy and psychometric rationales for why they are reasonable as reported. Evidence is provided that technical experts reviewed the appropriateness of the outcomes within the context of the assessment program, its intended uses, and the stated evaluation criteria presented in claim A.3.2.	In some cases this documentation may take the form of independent reliability studies, in other cases it will be results reported in a technical report. All reliability coefficients need not be reviewed, but rather a sample in order to get a good idea of the strengths and weaknesses of the program. While psychometric "rules of thumb" related to reliability are helpful for evaluating this claim, contextual factors of the testing program will influence the reasonableness of obtained reliability estimates (e.g., achieving an appropriate balance between validity (achieved through adequately broad, non-homogeneous content and DOK coverage) and reliability. The strength of the reliability results should be evaluated both in relation to the criteria specified in claim A.3.2 and also using expert judgment. For a newly developed/ proposed assessment that has not yet been administered, any reliability indices calculated using field-test data should be provided. If assessments have not yet been developed or field-tested evidence summarizing research /work consulted to inform the calculation or evaluation of proposed indices can be provided including examples illustrating results consistent with that expected.
Secondary claim related to informing users of reliability	***Quality of Evidence***	
	Sufficiency Statements	***Comments***
D.1.2.	*Evidence that users are appropriately informed of the magnitude of error surrounding the reported scores is essential for supporting valid interpretations of scores and their associated reliabilities. The sufficiency/ quality of the evidence presented in relation to claim D.1.2, therefore, should be taken into consideration when evaluating the claims associated with A.3, and when making a final, holistic determination regarding the strength of evidence presented in support of this criterion.*	

CCSSO Criterion A.4

The evaluation of evidence associated with A.4 involves judging the degree to which the provided documentation can support that assessments are designed and implemented to provide for valid and consistent score interpretations within and across years. The primary claims related to this criterion are divided into two main sections:

1) Assessment Development: claims A.4.1-A.4.5 evaluate the quality of evidence related to the item and test form development and review procedures.
2) Scaling and Equating: claims A.4.6-A.4.11 evaluate the quality of evidence related to the scaling and equating (or linking) procedures.

Additionally, because the validity and consistency of score interpretations depends also on the standardization of assessment delivery procedures and accessibility, two secondary sets of claims from criteria E.1 and A.5 are appended to support a holistic judgment regarding criterion A.4.

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.

Relevant Joint Standards (2014): 12.3, 12.8, 4.8, 4.9, 4.10, 12.6, 4.3, 4.4, 4.8, 4.11, 4.12, 2.15, 12.5, 4.18, 4.19, 4.20, 4.21, 5.2, 5.6, 5.7, 5.12, 5.13, 5.14, 5.15, 5.16

Primary claims related to assessment development	Quality of Evidence	
	Sufficiency Statements	**Comments**
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.	The construct or content domain of interest is clearly articulated. Specifically, those content standards deemed eligible for assessment are clearly identified so that there is no confusion regarding which KSAs will/will not be assessed. The type of evidence expected from students relative to each content standard is clearly defined so that content standards are not interpreted/ operationalized differently across phases of item development, or by different content developers. PLDs (even if preliminary) developed with the intent of supporting, item development (as discussed in A.1.4) are clearly incorporated into the item development process. Item development specifications and task models include enough detail to support consistency in the presentation, format, and degree of scaffolding observed in items and associated stimuli across forms. Evidence indicates that the item development specifications are produced by qualified personnel and reviewed for clarity and quality.	Articulating the construct or content domain that is the focus of assessments serves to reduce construct irrelevant factors that influence the consistency of score meanings across forms and years. The level of granularity at which expected evidence should be detailed will vary depending on the breadth and depth of the standard, and the types of items that are eligible for assessment (CR, SR, Technology-Enhanced, etc.). If a CAT engine is to be used, item development specifications include details related to how content and skill characteristics required by the items should be coded to support the requirements of the CAT algorithm and provide for the selection/administration of appropriate sets of items.

	A well-defined process is in place to support the maintenance and revision of item development specifications within and across years. The process should articulate the "owner" of the specifications, who is eligible to make modifications, and the process by which revisions to the document are suggested, evaluated and implemented. Item writer training materials include a discussion around the purpose and intended uses of assessment results, and a detailed description of the intended construct/content domain including expectations for cognitive demand. Issues of instructional sensitivity and fairness are appropriately addressed. Evidence indicates that procedures are in place to examine: 1) the effectiveness of item writing training procedures and 2) the impact associated with changes to these materials from one year to the next.	
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 process is in place whereby qualified content and accessibility experts review all newly developed test items for alignment to the standards[9] and adherence to the item development specifications. Clarity of items is reviewed to minimize construct-irrelevant variance. Items developed to support both online and paper-based delivery are reviewed for content/ skill based comparability in light of the format in which they will be presented. Items developed to support online and/or paper-based delivery are reviewed in the mode of delivery (e.g., online items are reviewed on computer). Any significant modifications in items from one delivery mechanism to another are acknowledged through treating those variations as different items in the review process. Any items presented in only one mode (or a subset of modes) of administration are reviewed to ensure that systematic differences in coverage are not presented across modes. Items go through a comprehensive bias/sensitivity review to make sure they are appropriate and fair for all relevant sub-groups and adhere to the principles of universal design.	Content experts should be are appropriately credentialed in their area of expertise and have experience in the grade range and developmental level of the students for whom the items are prepared. Comparability may be determined by content expert review or using cognitive labs. To ensure that reviews account for key accessibility concerns, the bias/sensitivity committee should have a good understanding of the intended test taking population and range of universal and selective accessibility features made available. *For assessments developed to be used in multiple states, procedures must be in place to ensure items will be appropriate and fair for all students regardless of the state in which they reside.*

[9] For a comprehensive evaluation of item and test alignment, please refer to the companion Test Content methodology.

	Content/bias sensitivity training materials include a discussion of the purpose and intended uses of assessment results, and a detailed description of the constructs to be measured by the assessment. Evidence of the adequacy and effectiveness of all reviewer trainings is reported. Checklists, guidelines and other reference materials are provided to reviewers to support them during their review. The qualifications, relevant experiences, and demographic characteristics of the reviewers are well documented. Reviewers should include, at a minimum, grade-level content experts, subgroup representatives/advocates, and accessibility experts for both ELL and SWD.	
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.	If a pilot testing sample of students is used, representativeness of the sample to the target population is documented. Subgroups of students are adequately represented, over represented, or weighted in analyses, as necessary, in the pilot testing of items. The quality and size of the pilot test activity account for the manner in which pilot test data will be used (e.g., solely to review item quality, support the estimation of item parameters used to support CAT testing). The psychometric quality of items is reviewed for difficulty, discrimination, fit, and differential item functioning (across sub-groups, mode of administration, and accommodations) to determine appropriateness for operational use for each purpose specified in A.2.1. Quality should be reviewed during pilot testing as well as after each operational administration. Flagging rules and/or evaluation criteria for poorly performing items needing careful, additional review are described in conjunction with a defensible rationale and detailed next steps. If DIF is detected, the actions taken to review, revise, and/or drop items from the item pool (or an operational form) are detailed and justified. Procedures allow for the re-evaluation of item "quality" definitions or flagging rules on an annual basis in response to changing contextual factors (e.g. level of implementation of the CCR standards and opportunity to learn the assessed content). If items are to be administered in multiple modes (e.g., paper and pencil and computer) or across	New curriculum/standards implementation will likely be a mitigating factor in reviewing, revising and evaluating items and the processes used in their development. This is especially problematic if schools, districts and states vary substantially in teacher preparedness and other resources needed to implement a new curriculum. The review of technical indices of item quality should not ignore the impact of gradual and varied implementation of new and challenging curriculum. Reviewers should examine items relative to the target they are designed to assess and consider whether pilot data may reflect inadequate opportunity for teachers to implement a new curriculum target. For example, items deemed appropriate via pilot testing may be flagged for psychometric reasons after operational testing. This is a common issue with DIF (particularly if there were not enough people in the subgroup during pretesting) but does come up for other reasons as well from time to time. If items that exhibit DIF and for which the bias-sensitivity reviewers have a plausible explanation for (i.e., a construct irrelevant factor) are dropped from the item pool, the process must ensure the coverage of the specified test content is not compromised. Qualitative evidence may include results from cognitive labs, focus groups or expert judgment. Sources of evidence could include the following, among others: • Detailed summaries of the executed pilot testing procedures.

	different platforms (e.g., tablet versus laptop), procedures are in place to test items in each mode or with each platform and test for comparability. Qualitative evidence is provided to support claims that new or innovative items (e.g., in terms of content, mode of responding, cognitive/physical/verbal requirements) are fair for all students and address the intended construct. Evidence is provided indicating that item pilot testing, evaluation, and review procedures were implemented as specified.	• A sample of items and along with indicators of their psychometric quality and fairness such as difficulty, discrimination, DIF, and item parameter consistency over time (e.g., data review reports). • Paper-pencil/Online comparability analyses for individual items. *For assessments developed to be used in multiple states, item pre-testing sampling and analysis plans should take into account the likely differing stages of implementation of new standards/curricula across states. Testing programs can deal with this in a number of ways including conducting DIF analyses or separate item calibrations.*
A.4.4. Test specifications clearly indicate how equivalent scores will be obtained across operational test forms within and across years.	Test specifications clearly articulate the level of equivalence that is considered adequate and how that will be determined from content (KSAs), test design, and statistical perspectives. For example, test specifications clearly outline the content and statistical rules (targets) underlying the composition of operational test forms (either fixed, or those resulting from CAT), when and how they were established, their rationale, and what (if any) degree of deviation is acceptable across forms and years. Procedures are in place to evaluate the appropriateness of fixed form or CAT test specifications from year to year. **Fixed Forms:** General Test Specifications should include: overall length and duration requirements for speeded tests, length requirements and typical and high-end expected durations for non-speeded tests, and details related to how items should be presented and ordered within and between test sections. Content-based targets include overall test length minimums and maximums, in addition to the expected representation of different standards, objectives, item types (MC, CR) and levels of cognitive complexity on a given form. Statistical targets are appropriately defined (e.g., classical or IRT) and clearly account for the manner in which scores are to be reported and used as defined within assessment development documentation. Statistical targets are established for the test overall as well as each reportable category/subgroup.	The rationale underlying the provided specifications should be consistent with the goals of the assessment and support the intended purposes and inferences. Test assembly rules that are based, in part, on pilot data should be re-evaluated after several years of program implementation. To ensure consistency in score meaning, the burden on test takers (from a time, difficulty standpoint) should be as consistent as possible from year to year. Content specifications for ELA specifically, should include rules related to the overall passage length, (e.g., informational/literary.), complexity, etc. to be represented on a given form. Specifications for math should detail expected level of precision as well as allowable supports (e.g., calculators, scratch paper, etc.). Statistical targets should be based upon the scale that will be used to inform, scaling, equating and reporting. If using IRT, then primary statistical targets should be IRT-based (TIF, TCC, etc.) and should be tied directly to the decisions being made on the test and subscore results. If for example, results are intended to support inferences regarding CCR, then targets should necessitate more precision (smaller errors) around those cuts. While it is typically recommended that the equating/linking set be proportionally representative of the total test in terms of content, when IRT based equating is used this

	Statistical targets include expected item difficulty (and range of difficulty), discrimination (if applicable), expected degree of measurement precision along the range of the reportable scale (e.g., test information function), and (when applicable) the raw score associated with particular scaled score cuts over forms. Target levels of measurement precision (and allowable variation) account for key inferences (CCR) or decisions that will be made in light of obtained scores or performance levels (mastery, graduation, etc.). If a linking/equating set is necessary, specifications and an associated rationale are provided around the size, location and statistical requirements underlying the selection of items for this set. The linking set should be of appropriate size/representation given the size of the assessment, the mathematical model being used (IRT vs. Classical) and whether items in the set are considered internal or external for purposes of scoring. **Adaptive Forms:** If a CAT or multistage testing is to be used, formal test assembly specifications are provided that show statistical targets, content and other related constraints, and the rules and associated rationales for the selection and administration of test items by the CAT/blocking algorithm (at item, testlet, and test level) including guidelines for determining starting points, termination conditions, and details related to exposure control. Where applicable, item bank inventory (supply) is reported relative to the test specifications (demands).	requirement is not as important as utilizing items that reflect strong psychometric properties (e.g., which provide more information and are less likely to be "unstable").
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.	Procedures are in place to ensure assessments (or for CAT, instantiations of an assessment) are reviewed by representatives with the appropriate level of content and psychometric expertise to ensure adherence to specifications. The criteria against which the reasonableness of a proposed form (or set of CAT forms) is to be evaluated, and how those criteria are prioritized, are clearly articulated. Evaluation criteria are reasonable and sufficient given the manner in which results are to be analyzed reported and used.	Evidence of appropriate review procedures may include the following, among others: • Detailed summaries of the executed item and test development processes and its results in terms of meeting the targets specified for information. • A sample of items and along with indicators of their psychometric quality and fairness such as difficulty, discrimination, DIF, and item parameter consistency over time. • A sample of test forms and along with indicators of their psychometric quality and fairness such as test information, test DIF, and difficulty.

	Detailed documentation is maintained throughout each iteration of the assessment (or CAT algorithm) development, review, and revision process. Procedures are in place for dealing with deviations from specified targets, or conditions under which targets cannot be met (e.g., constrained bank, etc.), so that all parties are made aware and understand potential implications. An adequate number of simulated (or real sample) tests from CATs are produced at all regions of the score scale and reviewed by content experts to ensure representation of content and cognitive processes. The evaluation of simulated CAT forms takes into account the size/breadth of the item bank in conjunction with exposure control specifications (i.e., to ensure that there are enough simulations to account for these factors). Statistics representing the entire scale should be produced and interpreted, to assure that systematic problems in one region of the scale cannot be washed out by counteracting problems in another region or by one or more non-problematic regions representing many students. Evidence is provided indicating that item and test form development and review procedures occurred according to plan. Any deviations in the planned procedures are accompanied with rationales.	
Primary claims related to scaling and equating	**Quality of Evidence**	
	Sufficiency Statements	**Comments**
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.	The properties of the reportable scale facilitate the use and interpretation of results as intended, and mitigate misinterpretations. a. The range and spread of the reportable scale provide for an appropriate floor and ceiling given the range of achievement expected (over the first few years) and intended uses of assessment results. The ceiling needs to account for improved performance once CCSS is implemented fully over the years. b. Key inferences (normative, predictive, criterion referenced) are highlighted or supported through properties of the scale. c. Supports are provided to minimize the misinterpretation of scaled score results due to factors such as a resemblance to other common or known scales.	The meaning of scale scores may be related to general proficiency standards, anchored in specific content and skill associated with different scale score levels, and/or norms for one or more specified populations of students. Scale construction should be consistent with the meaning stated or implied in score reporting. For example (with respect to bullet a.), since results are likely to be used to support growth or value-added models for accountability, the scale properties necessary to support these uses (as defined in research or through consultation with technical experts) should be incorporated into the design of the reportable scale.

	d. The connections between key positions on the scale score (e.g., cut scores) and the content of the items around that position are easily established and transparent. The reporting scale is designed in light of feedback from technical experts and intended users (e.g., teachers and parents) to ensure it supports the intended inferences and does not lead to consistent misinterpretation.	For example (with respect to bullet b.), consistent interpretations related to being on-track for CCR may be facilitated by associating this cut with a common scale score value across grades. For example, (with respect to bullet c.) the use of a 0-100 scale is typically discouraged because it may inadvertently provide for percent correct or percentile rank inferences.
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 coherent rationale is provided for the procedures selected/defined to support scoring and scaling of student results within and across grades, such as the choice of IRT model used to calibrate the items and transform the raw scores to a theta scale. Psychometric experts are involved in the review and approval of any studies conducted to evaluate the appropriateness, fairness and reliability of different scoring and scaling procedures prior to selection and implementation. Procedures for transforming raw scores, true scores, theta estimates or other estimates of student performance to the reportable scale metric are detailed enough to ensure accurate and consistent application across forms and occasions. - Decision rules are articulated for establishing HOSS/LOSS. - Rounding rules applied to translate score to the reportable metric are clearly specified. - If grade-level scales are used, procedures used to establish each scale and promote accurate interpretations across scales are described. There is documentation to describe the appropriate computation, reporting, and interpretation of scores for students with modifications or other non-standard administration of the test. Programs that attempt to maintain a common scale over time conduct periodic checks of the stability of the scale on which scores are reported. Technical experts were involved in the specification and/or review of the reportable scale.	For example, if a vertical scale is used evidence should be provided that studies were conducted and subsequently evaluated by qualified technical experts to determine the appropriateness of that scale. For example, if the assessment is intended to be a CAT or provide for a pre-equating design, Item Response Theory procedures must be used. It is often the case that item parameter estimates used to support the scaling of paper-based, accommodated forms (Braille) are based upon performance in the total population. If this approach is employed, evidence should be provided to support the fairness and reliability of these procedures.

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.	The process used to develop, review, monitor, and revise the scoring rubrics for accuracy, stability, clarity and fairness is clearly documented. The resulting scoring rubrics are clear and can be consistently applied by raters (i.e., provide for reliable scores). Scoring procedures are clear, comprehensive, and consistent with best practices. Scoring procedures are carefully monitored to assure that they are uniformly applied. Scoring rubrics are piloted before operational use, so that they can be modified as appropriate. The expected level of scorer agreement and accuracy (both instantaneous and longitudinal) is documented with an accompanying rationale. There are clear procedures in place for qualifying scorers and monitoring their performance throughout the scoring window. Audit and quality-check procedures (e.g. read-behinds, anchor sets) examine rater agreement and competence. Procedures are in place to prevent, detect, and, if needed, account for scorer drift over time. Documentation/evidence is provided which demonstrates that scoring procedures and results (especially as they relate to particular sub-groups) are not influenced by perceptions and predispositions of scorers.	If automated scoring procedures are used either alone or in conjunction with human scoring, comprehensive evidence is presented regarding the validation of the scoring algorithms to produce scores that accurately represent student achievement on the intended construct. For example, papers from prior administrations may be scored along with papers from the current administration, where prompts are repeated.
A.4.9. Linking and/or equating procedures are clearly specified, comprehensive, and demonstratively appropriate.	Documentation of equating procedures provides enough detail to allow for independent replication of all procedures and results. Documentation should include: equating design and method (e.g., pre-equating/post-equating, randomly equivalent groups, common item non-equivalent groups, etc.), specifications for sample used to estimate item parameters or score distributions, decision rules applied (e.g., related to the stability of linking items), procedures used to validate the quality of the equating results, and any methods used to update item parameters for inclusion in the item bank after operational administration (when applicable). A comprehensive and defensible rationale is provided for procedures used and their associated evaluation criteria.	Equating design should make sense in light of administration, scoring and reporting timeline and quality/representativeness of available data. For example, a pre-equating design is not appropriate if item level data is based upon an unmotivated field-test sample. In such cases post-equating verification is necessary. For example, plans to use samples of examinees or only some item types or content should be supported by research showing that such an approach will provide results equivalent to those based on all students and all item types/formats. It is common to equate writing tests based on multiple choice items only because of difficulties with repeating prompts. It is also common to drop items from an equating anchor because of aberrant trend information. Such procedures

	Procedures are in place to calculate and evaluate the standard error of equating at different points along the score scale continuum and over more than a single linking procedure. Criteria for evaluating standard errors are provided in conjunction with actions that will be taken if issues arise. A comprehensive replication or validation process is in place to assure the accuracy of equating results. The process outlines, at least, the parties involved, process used to ensure independence, rules for evaluating the consistency of results between parties (e.g., degree of precision required, etc.), and procedures used to resolve any differences observed. The process defines where, exactly, in the equating process replication begins (e.g., after data cleansing, prior to calibration, etc.) and, consequently, the type of errors that may/may not be caught in the replication process.). Replicators independently write their equating code and/or implement equating procedures with only the description of the equating methodology, sources of data, and exclusion rules to assure that the descriptions are adequate to allow for replication without sharing code or other software commands. Procedures and documentation are in place to evaluate the feasibility/reasonableness of equating results (e.g., in light of historical data, quality of the data used to establish the baseline scale, etc.…), including: evaluation of trend lines, historical impact data, and raw score cuts. When appropriate, procedures are in place to evaluate the comparability of the equating across subgroups of sufficient size. Procedures are in place to monitor equating stability over time and detect scale drift (e.g., across multiple forms in a chain, across time with an adaptive pool) that might occur due to curriculum implementation.	should be documented along with procedures for ensuring consistency in content coverage of the reduced equating anchors. The scope of this replication will differ depending on the nature of the equating design and the mode of administration. • If pre-equating is used then replication involves the development of scoring tables in light of pre-equated item parameters or data. • If post-equating, replication typically involves all stages of the equating process, starting with the application of exclusion criteria and ending with the development of scoring tables. Replication by analysts independent of the organization responsible for equating is preferred over replication by an independent group within the same organization. Replications using both the same software and different software are preferred. External or independent reviewers should be tasked with critiquing those results and posing questions/concerns for follow-up.
A.4.10. The scaling and linking/equating procedures were followed as specified, and the results have been reviewed and accepted by technical experts.	Documentation is provided that indicates the scaling and equating/linking was conducted as intended. This documentation may include: • Detailed summaries of the executed procedures with relevant results included. • An independent analysis or QC report which verifies that all calculations were completed accurately.	The quality of the evidence presented will depend on the extent to which the experts are independent, and the amount of evidence reviewed. External reviewers are preferable, to internal reviewers. However, internal, independent reviewers are preferable over less transparent quality control procedures.

	• A summary report of the replication, including any errors found, clarifications to instructions needed, and deviations between original and replicate analyses. Evidence should include documentation of expert review and acceptance of the following: • Assumptions necessary to support the selected equating procedures (e.g., in sampling students or items) are reasonably confirmed by research studies. • The observed equating results. • Standard errors of equating across the achievement continuum, equating constants, properties of the linking/equating set -- including items dropped for lack of stability. The qualifications and experience of the experts is presented.	Any deviations in the planned procedures are accompanied with rationales and evidence that the validity of the performance level descriptors was not sacrificed. It is hoped that the observed equating results would show the following: test forms are not found to be unexpectedly difficult or easy; there are no large, unexplained changes in student performance over time. Results of procedural checks (on form difficulty, score trends overall and by group, etc.) do not indicate unreasonable unexpected and unexplained variances.
Secondary claims from E.1 related to assessment standardization	***Quality of Evidence***	
	Sufficiency Statements	***Comments***
E.1.1 – E.1.2	*Evidence related to the quality and standardization of the distribution and administration procedures may inform decisions about the consistency of score meanings within and across years. The sufficiency/quality of the evidence presented in relation to claims E.1.1 and E.1.2, therefore, should be taken into consideration when evaluating the claims associated with A.4, and when making a final, holistic determination regarding the strength of the evidence presented in support of this criterion.*	

CCSSO Criterion A.5[10]

Criterion A.5 is a special case in that all of the sub-criteria are evaluated in both the Test Content and the Test Characteristics methodologies. For example, the first two sub-criteria, related to universal design and the provision of appropriate accommodations/modifications, are addressed in the Test Content methodology from a construct/content perspective (i.e., using item development/review documentation and samples of test items), and then again in the Test Characteristics methodology from a reliability/validity perspective (i.e., using information and data gained before, during and as a result of test administration). The final two bullets relate to the evaluation of evidence demonstrating that accessibility features provide for reliable scores and valid inferences specifically for English learners and students with disabilities. These issues are considered from a process/documentation stand point in the Test Content methodology, but addressed from a technical standpoint in the Test Characteristics methodology. Additionally, 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.

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[11] 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.**

Relevant Joint Standards (2014): 3.6, 3.7, 3.8, 3.9, 3.10, 3.11, 3.12, 3.13, 3.14,3.15, 3.17

Primary claims relating to the test characteristics associated with accessibility	Quality of Evidence	
	Sufficiency Statements	**Comments**
A.5.1. The testing user interface[12] and item format does not introduce construct-irrelevant variance that impedes student performance.	Evidence is provided which shows that: • The general population of students, including all relevant subgroups, is able to navigate and access test content, and is able to efficiently produce intended responses. Required tools, such as calculators, measuring devices, and equation editors, are intuitive and easy to use. • Accommodated[13] test items and accessibility features permit students to demonstrate their knowledge and abilities and do not contain features that prevent them from accessing the content of the items. • Scores of students who receive accommodations/choose accessibility features are not unduly influenced by construct irrelevant variance related to administration and scoring of accommodated test forms or items. Evidence may include documentation that scores assigned to students representing subgroups of	During item development, evidence would include use of guidelines for creating alternate representations of content, including audio/read aloud, ASL, braille/tactile, and/or translated versions of item content. These alternate versions of content would then be reviewed by experts for adherence to the guidelines and quality. Additional evidence may include documentation of the approval of trans-adapted forms by an expert reviewer and results/feedback from cognitive labs in which students are asked to translate/interpret the intent of each item. The choice to use accessibility features often varies by item. In order to establish that these access tools do not unduly influence performance, the use of any tool would need to be collected on an item-by-item basis and rules for classifying tool user vs non-user are needed. Plans should be in place to collect this and similar types of

[10] While both the Test Characteristics and Test Content methodologies contribute to the evaluation of Criterion A.5, the development of the respective methodologies occurred independently of one another and therefore, the claim labels (e.g., A.5.1) are too independent and do not carry any meaning within the Test Content methodology.

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

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

[13] The definition of accommodations, for the purposes of this evaluation, includes traditional Individualized Education Plan (IEP)-dictated accommodations (e.g., extra test time, separate room) as well as student self-selected accessibility tools (e.g., electronic read-aloud). Evidence should be provided to support the use of all available accommodations.

	the population have similar levels of reliability and validity when compared to those associated with that total population. For example, students taking certain accommodations have similar levels of reliability as those associated with students taking non-accommodated forms. There is a clear plan in place for evaluating the effect of accessibility features on score validity (e.g., a validity argument for accommodations, online features; documentation of development; documentation of training; documentation of use; documentation of effect). Guidelines have been established to ensure that the use of interpreters, when necessary, does not introduce construct irrelevant variance or influence the nature of the construct being assessed.	information in order to document the differences between items and better understand the implications for validity. Interpreters should follow standardized procedures, and be sufficiently fluent in the language and content of the test and the examinee's native language and culture to translate the test and related materials and appropriately represent the examinee's test responses, as necessary.
A.5.2. Students are matched with appropriate accommodations/ accessibility features.	Clear and standardized guidelines are in place to support the assignment of students to appropriate accommodations. Guidelines include a variety of options and clearly indicate conditions under which the use of particular accommodations contributes to score validity. Guidelines are in accordance with all state laws and polices concerning the access to and assignment of accommodations. Tools are provided and readily available to assist with decision-making (e.g., decision-making trees, FAQs). Evidence is provided that instructors and/or IEP teams are able to appropriately assign students who are English learners and students with disabilities (SWDs) to accommodations. This evidence may take the form of feedback from educators that they understand the rules related to assignment to accommodate forms and/or that they are consistently assigned across educators. When students are allowed to self-select from a set of accessibility features, evidence is provided to show that students have adequate training and familiarity with the options to make informed decisions. For example, this evidence can be derived from cognitive labs or accounts from pilot testing. As part of administration procedures, on-going evidence is collected that monitors the appropriateness of access to and implementation of accommodations and accessibility features for all student groups. Procedures are in place so that study results inform future training and monitoring needs.	Much of the effectiveness of an accommodation is determined by whether the right students get the accommodation (and wrong students do not), and how the students use the accommodation. In order to demonstrate evidence related to this claim, the testing program will likely need to have procedures in place for associating student scores with the types of accommodation(s) or modification(s) that were used for the test. *For assessments developed to be used in multiple states, the testing program should have procedures in place for ensuring that the guidelines provided to schools and districts are in accordance with the all relevant individual state laws and policies concerning accommodations.*

A.5.3 Score reliability is appropriately estimated and evaluated for English learners and students with disabilities (SWD).	Appropriate psychometric procedures are in place to estimate and evaluate the reliability of assessment results for English learners and SWD. If the reliability evaluation criteria differ from those presented in relation to criteria A.3., a clear rationale is provided.	For example: • Different language and disability groups have different needs, and so where feasible (i.e. sample size permitting), analyses should be separated by disability grouping (e.g., communication disabilities, cognitive disabilities, physical disabilities, etc.). Reliability could be analyzed by accommodation, comparing those who received it and those who didn't, regardless of disability. Either way, analyses should support that no one disability group is disadvantaged and that the support-use provides and "equal playing field."
A.5.4 Validity evidence supports the intended use and interpretation of scores for English learners and students with disabilities (SWD).	Evidence is available to support the validity of the intended score interpretations (e.g., CCR claims) for all students. That is, evidence collected to validate the use of assessment results as an indicator of college and career readiness is provided for English learners and students with disabilities. A rationale, supported with empirical evidence, demonstrates that when each accommodation/ accessibility feature is used as recommended (either alone or in combination), the resulting scores are comparable to non-accommodated tests. This type of comparability evidence includes results of analysis investigating differential item functioning (DIF), speededness, and factor structures. If credible research indicates that scores do not have comparable meaning across subgroups, appropriate cautionary statements are provided. Evidence is collected showing that each accommodation/accessibility feature provides the support intended. Evidence suggests that the accommodation provides for gains in the intended populations. In the case when accommodations or accessibility features are used together, evidence shows their interaction of the two results in only intended or desirable consequences. Studies are planned/completed to evaluate the alignment across IEP-determined instructional accommodations and IEP-determined assessment accommodations.	Different language and disability groups have different needs, and so where feasible, analyses should be separated by disability grouping (e.g., communication disabilities, cognitive disabilities, physical disabilities, etc.). Some assessment programs offer dozens of accommodations/access features. Since there are possible interaction effects, each combination should be validated. As the number of possible accommodations/access features increase, the number of possible combinations also increases. In the case where thousands of possible interactions exist, a defensible sampling scheme, which includes the most frequently occurring combinations, should be clearly articulated along with a rationale. In some cases, sample sizes of student groups (especially for some languages and/or more rare disabilities) will be small. Reasonable attempts should be made to understand the impact of accommodations/accessibility features on score validity for these groups (e.g., case studies).

Secondary claims from A.4 related to item review	*Quality of Evidence*	
	Sufficiency Statements	*Comments*
A.4.2-A.4.3	*Evidence related to the processes used to design, develop, and review items for bias/fairness may inform decisions about the accessibility and appropriateness of test items for students with disabilities and English learners. The sufficiency/quality of the evidence presented in relation to claims A.4.2 and A.4.3, therefore, should be taken into consideration when evaluating the claims associated with A.5, and when making a final, holistic determination regarding the strength of the evidence presented in support of this criterion.*	

CCSSO Criterion A.7

The evaluation of evidence associated with A.7 involves judging the degree to which the provided documentation can support that the assessments meet all federal and state requirements for student privacy and all data is readily accessible by the state. The primary claims related to this criterion are divided into two main sections:

1) Student Privacy: claims A.7.1-A.7.3 evaluate the quality of evidence related to the procedures in place for ensuring student privacy and data security.
2) Data Access: claims A.5.4-A.5.6 evaluate the quality of evidence related to assurance that the state and other relevant stakeholders have appropriate access to data in order to meet their needs and carry out their respective responsibilities.

In addition to the primary claims associated with this criterion (A.7.1-A.7.6) secondary claims related to criterion E.1 are appended. Claims E.1.3-E.1.6 call for evidence related to the security of testing materials, which is essential for ensuring student privacy and, more peripherally, state access. The strength of the evidence provided for claims E.1.3-E.1.6 should be considered when making a holistic judgment regarding criterion A.7.

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.*

A.7.2. Comprehensive procedures are in place to protect personally identifiable information (PII) from unauthorized access or use.	Documentation specifies all personally identifiable information that is collected as part of the assessment process (so it is clear to all involved which data does/does not fall into this category), how it will be used, and why it is necessary to support the assessment program. PII is only stored if it has a clear, specific purpose. Documentation includes the safeguards in place to protect against the loss or unauthorized access, use and disclosure of secure test materials and score reports that include or can be associated with PII. Access to materials containing, or that can be linked back to, PII is only provided to authorized persons, and then only on a "need-to-know" basis. The procedures and technology used to monitor and maintain the security of PII throughout the life of an assessment program are clearly defined. Mechanisms are in place to monitor for security breaches that may compromise PII.	State and local laws concerning PII will vary within each testing program. Security plans will either need to reflect this diversity or organize all procedures to meet the most stringent data protection regulations.
A.7.3. Procedures are in place to ensure all data is managed securely.	Procedures for stewardship and access to different types of data are coherent and compatible across the entire assessment program. Evidence is provided that documented procedures related to the secure management of sensitive data are known and enforced by all who have access to such data. This can be achieved through audits, trainings, and confidentiality agreements within the testing program and its contractors. Data management systems must have enforceable access policy rules in place for not only source data but also for extracts or any interim files or auxiliary data images created A formal response plan is in place for appropriately handling a breach of confidentiality or unauthorized access to personally identifiable information. The response plan includes root cause analysis and steps to ensure the event will not happen again in the future.	*If a multi-state assessment is administered, scored and/or reported by a common vendor across states, many people may have access to the test data at different points of the development, administration, scoring and reporting phases of the assessments. The testing program is ultimately responsible for monitoring compliance with program-specified security standards to maintain the validity of the test score interpretations and uses. The quality of evidence related to this claim will rely on the extent to which the testing program has procedures in place to ensure data security.* *For assessments developed to be used in multiple states, if the assessment is not administered, scored and/or reported by a common vendor across states, this claim can only be partially evaluated if common guidelines/rules are provided to ensure appropriate secure data management/security procedures are used. However, broader inferences related to compliance (consistent with the statements provided to the left) will need to be evaluated on a state-by-state basis.*

Primary claims related to data access	Quality of Evidence	
	Sufficiency Statements	**Comments**
A.7.4. An assurance is provided of state ownership of all required data[14] reflecting compliance with state laws.*	The process by which the ownership, use, and transfer of any data related to or resulting from the assessment (i.e., between the state and test vendor) are clearly outlined. The process outlines when these discussions will occur and what elements are necessary for consideration. Caveats or exceptions related to rules for ownership are clearly defined. Intended ownership of assessment data (i.e., student performance data) or materials (e.g., test items, item banks, item-level performance data, etc....) at different stages in the contract (including contract end) should be explicitly outlined by the test vendor in conjunction with a clearly articulated rationale for this decision. The ownership assurance should include provisions related to 1) who has complete access to the data, and 2) what the vendor is allowed to do with the data even under state ownership or co-ownership. States should never be denied access to data necessary for replication and/or quality control due to vendor policies related to data ownership (at least during the duration of a contract).	Rules related to ownership of student data will vary depending on a variety of factors, including the nature of the assessment (e.g., newly developed vs off-the shelf) and who it was developed to serve (e.g., a state vs. a consortium). Specifications related to ownership must be clearly articulated by the test vendor so that they can be evaluated by the state relative to state laws and requirements. *For assessments developed to be used in multiple states, the program's policy must either hold for the most stringent state's laws about data ownership, or, develop differentiated policies by state.* *Note: In some cases, ownership of data by the state may not be necessary or possible; however, state ownership must be the vendor's policy in order to receive a rating of "Good" or better on Criterion A.7.
A.7.5. Procedures and timelines are in place to ensure a state[15] is provided with all data necessary to support desired analyses (e.g., forensics, quality control, accountability calculations) in a timely and useable fashion.	A comprehensive process is outlined for collecting, documenting and adhering to state specifications related to the need for underlying data including: required format, schedule, completeness (e.g., total population vs. a sample) and quality (e.g., pre- or post-edit phase). Within the bounds of contractual agreements between states and vendors, states have "on demand" access to source data for test forms, items, student demographic data, response data, and all test scores. Standardized, portable formats for data sharing "on demand" are in place. Also, the procedures and control files/data used to generate the data are accessible (e.g., archived response files and controls files used for scoring). Sample files and associated file layouts are provided to the state prior to use to allow for the state to evaluate their utility (in content and format) in supporting intended analyses.	Depending upon the requirements of the state, this may require transfer prior to the reporting of any test results. The schedule for the transfer of the files should be either specified in the contract or within another legally binding agreement. *For assessments developed to be used in multiple states, participating states should be provided the file layouts prior to data transfer to allow for feedback and comment. It is the testing program's responsibility to ensure the data files will be usable by all states it serves.* *For assessments developed to be used in multiple states, if the assessment is not administered, scored and/or reported by a common vendor across states, this claim can be partially evaluated if common guidelines or materials are provided to all states to help to support the articulation of data needs (for analyses) and appropriate timelines. However, broader inferences related to compliance (consistent with the statements provided to the left) will need to be made by evaluating evidence on a state-by-state basis.*

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

[15] 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.

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.	The means by which required data and test results will be securely transferred between parties are clearly outlined (e.g., secure FTP; encryption, etc.). Procedures for stewardship and access to different types of data are coherent and compatible across the entire assessment program. Detailed specifications outline which employees are eligible to access and authorize the transfer of secure data. The contract between the state and the test publisher specify the policies and procedures for the transfer of any data, including: acceptable formats, metadata, data dictionaries, who can request the transfer, quality control steps in the transfer, and schedules for carrying out the transfer. Upon completion of the contract, procedures for storing or destroying data by the vendor are clearly articulated along with a rationale. If any data become the sole property of the vendor upon completion of the contract, states should have continued access to the data for replication/ quality control procedures.	
Secondary claims from E.1 related to security of test materials	***Quality of Evidence***	
	Sufficiency Statements	***Comments***
E.1.3-E.1.6	*Evidence related to the procedures for ensuring the security of assessment materials may inform decisions around the appropriateness of data privacy and data access. The sufficiency/quality of the evidence presented in relation to claims E.1.3-E.1.6, therefore, should be taken into consideration when evaluating the claims associated with A7, and when making a final, holistic determination regarding the strength of evidence presented in support of this criterion.*	

CCSSO Criterion D.1

The evaluation of evidence associated with D.1 involves judging the degree to which the content and format of score reports support the intended uses and interpretations of the assessment scores.

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.		
Standards: 6.10, 1.3, 1.14, 1.15, 12.11, 12.18, 5.1, 5.2		
Primary claims related to score report content and format	**Quality of Evidence**	
	Sufficiency Statements	**Comments**
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.	The reportable variables, categories, or sub-scores reflected on score reports are appropriate given the emphases reflected in assessment design documentation such as test blueprints, specifications, the theory of action, and scaling procedures. Score reports provide an accurate representation of the knowledge and skills emphasized by the test. If the theory of action emphasizes the provision of a certain type of data or information as important to achieving the goals of assessment (e.g., growth to standards) this emphasis should be and appropriately represented on the score reports. The reporting structures are defensible given the psychometrics characteristic of the test (e.g., number of items/score points at the test and sub-score level, reliability). Any reported score or associated standard error of measurement should be consistent with the applied scaling procedures (e.g., classical test theory vs. IRT), and be supported by the design of the assessment (e.g., the number of items and associated precision). Both numerical and graphical representations of student achievement on each report appropriately reflect the purpose of the report and the intended user.	Samples of all types of score reports that are generated to present to results of the assessment should be provided for review (e.g., different reports for different audiences, multiple grade levels if reports vary by grade).
D.1.2. Score reports support inferences regarding student achievement relative to key content and performance standards.	Score reports go beyond the reporting of scale scores, and provide for results that are clearly aligned with performance standards (i.e., cut scores) and defined by PLDs. Documentation is provided that indicates the intent of each score report and its primary audience. Results are disaggregated and reported in a manner that allows for the evaluation of student achievement relative to key content standards (or clusters of standards) to the degree that such	Reports should facilitate the use of results by stakeholders as intended during assessment design. The grain-level at which results should be presented on a given report depends on the purpose of that report and the manner in which scores are to be used. For example, if the primary purpose of a report is to help educators understand how their students are performing relative to CCR performance standards, student performance at the standard/objective level may not be necessary. However, if the goal is to provide educators with feedback that informs program improvement decisions, instruction and/

	subscores are supported by the assessment design. 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. 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). 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. Evidence is provided which shows that feedback is gathered from a representative sample of target end- users and is used to evaluate the degree to which reports are clear, user friendly and will be used and interpreted in the manners intended (e.g., empirical studies such as cognitive labs showing that users can accurately interpret the results to answer questions). Feedback is used to revise score reports as appropriate. Such reviews occur multiple times in the design/development process to take advantage of the resulting feedback. Documented evidence should be provided (e.g., technical reports summarizing usability studies).	or remediation the report should aggregate results at the finest-grain level at which the results are supported by reliability/precision and validity evidence. If the assessment measures only a subset of the total universe of standards, it should be made clear which standards are/are not assessed. The assessment program articulates the conditions under which they can expect the score reports to be correctly interpreted. For example, if professional development is required, the assumption that all users have received the requisite training is clearly stated. *For assessments developed to be used in multiple states, if states are using different score reporting procedures and formats, this claim will need to be evaluated on a state-by-state basis.*

D.1.3. Score reports provide for valid inferences regarding career and college readiness, or on-track to CCR.	Reports indicate how students are progressing relative to CCR or on-track standards. Any representation of progress toward CCR is reported in conjunction with a clear description as to how readiness is defined. The manner in which "progress" on the CCR continuum is represented is consistent with, or reasonable given, the way in which the test is scored (e.g., vertical scale) and CCR standards are defined. The ways in which representations of progress toward CCR are/are not to be interpreted are clearly articulated on score reports. Evidence about the effectiveness of the score reports in terms of: (a) ease of interpretation, (b) accuracy of the interpretations; and (c) reducing potential misuses is provided. This evidence could include documentation that shows that feedback is gathered from individuals or focus groups to evaluate users' understanding of how to interpret and use scores representing student progress, and used to revise score reports as appropriate.	For example: If grade-level CCR standards are defined in terms of the knowledge and skills necessary to be on-track to be CCR upon exit from high school as reflected in PLDs, progress may be defined in terms of a student's location on the reportable scale relative to this standard. If a vertical scale is applied – progress may be represented in terms of gains in student performance across grades, relative to that expected to be on-track to be CCR upon graduation. If the benchmark reflecting CCR or on-track to readiness is determined in light of particular probability of success on a criterion measure of interest (e.g., state test in next grade, readiness for credit-bearing first year college courses) progress may be similarly defined in terms of a probability of performance or success with respect to this predictive criterion variable (e.g., The student's 4th grade score is associated with a 85% probability of proficiency on the 5th grade test, which is higher than the on-track benchmark that was defined as having a 65% probability of proficiency). If growth scores are calculated for individual students, these scores must be clearly defined, and evidence of their validity, reliability, and fairness should be reported. *For assessments developed to be used in multiple states, if states are using different score reporting procedures and formats, this claim will need to be evaluated on a state-by-state basis.*

CCSSO Criterion D.2

The primary claims related to this criterion are divided into two main sections:

1) Access to Score Reports: claims D.2.1-D.2.1 judge the degree to which score reports are readily accessible to stakeholders and provide timely data consistent with the intended use.
2) Instructional Utility of Score Reports: claim D.2.3 evaluates the strength of the evidence to support the instructional value of score reports for providing useful, actionable data to students, parents, and teachers.

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.

Standards: 6.13, 12.19

Primary claims related to access to the score reports	Quality of Evidence	
	Sufficiency Statements	**Comments**
D.2.1. Directions for accessing and viewing score reports (when necessary) are broadly distributed and clear to end-users.	If reports are not directly distributed to users, adequate procedures are in place to ensure stakeholders access. Information regarding where and when to access score reports, as well as any security or password needs, are provided to stakeholders using multiple channels (e.g., web sites, newsletters, letters home, etc.) and in as timely a manner as possible. If online reporting systems allow for different "views" or the creation of custom reports, directions as to how to accomplish this are clearly articulated. Feedback from focus groups or other types of stakeholder engagement is considered when designing the online reporting system. User-customizable interfaces are scrutinized for potential misuse of the data or results (e.g., inappropriate aggregations when results fall below a certain count or user-specification of graphical displays that may over-emphasize small differences as important). Feedback gathered from individuals or focus groups is used to evaluate users' understanding of how to access and view and interpret score reports. Such feedback is used to revise the distribution process.	*For assessments developed to be used in multiple states, if states are using different score reporting procedures and formats, this claim will need to be evaluated on a state-by-state basis.*
D.2.2. Reporting timelines, procedures and technology provide for the dissemination of test results in a timely fashion.	The schedule for score reporting is well-articulated and publicly available. The timing for the release of score reports allows for test results to be used and interpreted as intended. (That is, the timeline clearly accounts for the needs and intended uses of different test results for different stakeholders). The procedures and technology in place to issue score reports /test results facilitate the use of	The timing and format and technology used to support the provision of test results (e.g., dynamic reporting, rolling reporting) should clearly account for the manner in which data is intended to be used (and by whom) at a given point in time. For example: if results are intended to support instructional planning, they should be available prior to the onset of a new school year. Similarly, if overall student scale scores are required to support teacher or school

	test results as intended/required by different stakeholder groups. Procedures include protocols for quickly responding to errors within score reports. The information regarding the error along with corrected score reports are distributed as soon as possible to all recipients to minimize misinterpretations of erroneous scores.	accountability, these scores may be required on an expedited schedule. In this case an initial "student summary report" that provides only this information in a readily consumable format may be appropriate. Similarly, for purposes of identifying appropriate remediation plans, students and higher education institutions may need a report as soon as possible after testing that tells them whether or not they achieved the cut-score necessary to be deemed ready for credit bearing work in college. This could be provided to the school or each student via e-mail or a simple PDF report. *For assessments developed to be used in multiple states, if states are using different score reporting procedures and formats, this claim will need to be evaluated on a state-by-state basis.*
Primary claims related to instructional utility of score reports	**Quality of Evidence**	
	Sufficiency Statements	**Comments**
D.2.3. The content and structure of score reports provide useful and actionable information for making instructional decisions.	Reports intended to be instructionally valuable are clearly identified, and the design process reflects a deliberate attempt to achieve this goal (i.e., in terms of both content and format of the report). The intended instructional value of each report is identified in advance and the features of the report align with that intent. The report design/development process provides evidence of the utility of the reports for making instructional decisions. Descriptions of development processes include results of studies (e.g., focus group studies) conducted to evaluate the degree to which reports provide instructional value to teachers, parents and students (of the type intended or expected). Feedback collected from teachers includes information related to the usefulness of reports and data for informing instructional practices. Feedback from parents/students may include that related to the utility of reports for informing decision making (e.g., course selection, remediation needs, etc.). Study samples are representative of the intended users (e.g., students, parents, and teachers). Scores are reported at the finest grain size possible as supported by reliability and validity evidence. The content and format of score reports facilitate the likelihood that scores will be used to inform instruction in the manner intended. Content refers not only to the scores presented, but also any interpretive text provided on reports to support the use/interpretation of those scores.	The types of instructional decisions supported by different score reports will vary – in general, and for different stakeholder groups. Feedback collected from stakeholders should be focused on determining what, if any, instructional value a given report provides relative to that intended. While the summative assessments do not have formative value, the instructional value for assessment results could include identifying students who need remediation, evaluating instruction for future improvement, and making curricular modifications for the future. Some reports may not be designed to support instruction at all, and therefore would not require such feedback. If score reports include recommendations for instruction intervention or are linked to recommended instructional plans or materials, rationales and evidence to support those recommendations should be provided. Not all testing programs will specify instructional decision-making as one of the intended or supported uses of the reported scores. However, in order to receive a rating of "Good" or better on Criterion D.2, instructional decision making must be one of the intended and supported uses of the reported scores. *For assessments developed to be used in multiple states, if states are using different score reporting procedures and formats, this claim will need to be evaluated on a state-by-state basis.*

CSSO Criterion E.1

The evaluation of evidence associated with E.1 involves judging the degree to which the provided documentation can support the standardization and security of the testing materials for the purpose of maintaining validity, fairness, and integrity of test results. The primary claims related to this criterion are divided into two main sections:

1) Standardization: claims E.1.1-E.1.2 evaluate the quality of evidence provided relating to assessment distribution and administration procedures and the adequacy of standardization across within and across years.
2) Security: claims E.1.3-E.1.9 evaluate the quality of evidence provided relating to test security procedures including the prevention and detection of security breaches to ensure the on-going integrity of the testing program.

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.

Standards: 6.1, 6.2, 6.4, 6.5, 12.16, 4.5, 4.15, 4.16, 6.6, 6.7, 9.21-9.22, 12.7

Primary claims related to standardization	Quality of Evidence	
	Sufficiency Statements	**Comments**
E.1.1. Test distribution and administration directions are clear and sufficiently scripted to provide for standardization.	For Computer-Based Testing (CBT), when equipment or software is likely unfamiliar to test takers, students should be given ample opportunity to practice logging in, navigating a sample test, and accessing online tools prior to the operational administration. For paper-based testing, any procedures that may be unfamiliar to test takers or administrators should be explained in advance, with opportunity for practice. Written instructions to examinees must make response expectations clear for all test takers and include ample practice items. Any materials used to support administration are carefully reviewed and pilot tested at least with a small sample of the intended audience or users. Feedback from sample is incorporated to improve standardization of test administration. Formal procedures are established for requesting and receiving accommodations, and documentary evidence shows that test takers have been informed of these procedures sufficiently in advance of administration. The directions and instructions for test form distribution and administration are clearly specified for each test administrative role (e.g., proctors, administrators, test director), including exception/incident handling and quality control procedures. For group testing and unless otherwise provided as part of a computerized test administration, directions to the proctors include a script to be read to test takers to ensure all students are given standardized instructions.	*For assessments developed to be used in multiple states, if states are using different vendors to support the administration, this claim will need to be evaluated on a state-by-state basis unless common templates, guidelines, scripts, directions or manuals related to assessment distribution and administration were generated as part of the test development process and are being used across all states.*

	Directions provided to test takers *within the context of the assessment* are of sufficient detail for test takers so that they respond to tasks in the manner intended by the test developer. Test Administration manuals describe permissible variations and exceptions from established standardization (e.g., accommodations for students with disabilities) requirements; an accompanying rationale for the allowable differences is included. There is a help desk available to answer procedural questions in a timely manner related to assessment administration during all test administration sessions. Technical help is available for all CBT administrations. Procedures are in place for monitoring if directions and administration procedures are followed, if only on a sampling basis, and in cases where they are not, the response is detailed. The response to irregularities includes reporting to the state when deviations from administration procedures appear intentional or nefarious and require additional investigation.	
E.1.2. Procedures for training and monitoring test administrators are effective and well documented.	Detailed specifications are provided to test administrators around the environment/setting within which tests must be administered including, but not necessarily limited to those related to: timing, location, student seating, access to personal electronic devices, rules related to leaving the testing environment (physical or virtual), rules for stopping, exception handling and, for CBT, procedures for disabling access to web browsing or other outside resources. Proctors are required to verify the identity and eligibility of all examinees and, when applicable, confirm assigned seating and spacing between seats for group testing. Any allowable tools/supports are made available to students and provided as part of the testing process (Scratch paper, calculators, rulers, etc.). Procedures are in place for monitoring that the specified conditions under which the tests are administered are followed and in cases where they are not, the response is detailed.	These specifications will sometimes discuss covering materials posted on walls that may be related to the content of test items. For example, when assessing a child's ability to identify adverbs, classroom posters referencing the parts of speech should be covered or removed. *For assessments developed to be used in multiple states, if states are using different vendors to support the administration, this claim will need to be evaluated on a state-by-state basis unless common templates, guidelines, scripts, directions or manuals related to assessment distribution and administration were generated as part of the test development process and are being used across all states.*

Primary claims related to security	Quality of Evidence	
	Sufficiency Statements	**Comments**
E.1.3. Comprehensive procedures are in place to ensure the security of assessment materials.	**Materials under state or vendor custody:** A clearly documented chain of custody exists throughout the item and test form development, review, and publishing cycle such that participants know who is responsible for ensuring the security of assessment materials (e.g., items, test forms, ancillaries) at any given time. The chain of custody is maintained once data has been returned to vendors for scoring. For non-electronic transfers, the chain of custody is tracked with sign-offs at every point of transfer. If item or test form development/review occurs online, items are stored on secure servers and there is a robust data management system in place to provide differentiated levels of access to different users with respect to specific data repositories, data tables and even specific records. Security procedures are followed at workshops (e.g., standard setting, item writing, content & bias reviews) to protect the integrity of the materials. Clear documentation outlines the security procedures related to scoring, hand scoring, data analysis, data transfer, and reporting. **Materials not under state or vendor custody:** Paper forms are delivered to local education agencies in a manner (e.g., shrink wrapped or with wafer seals) that prevents review prior to administration. When tests are with local education agencies, a chain of custody for testing materials is clearly articulated. Procedures clearly indicate each individuals' responsibility from the time of materials receipt to materials return, and includes a mechanism by which to record transfer of responsibility of sensitive materials, data, and information (e.g., sign in/sign out documents, bar code scanning, etc.…). A process is in place to ensure that when test materials transferred to local education agencies, all secure materials are accounted for both upon receipt and return. Procedures dictate how secure materials must be stored between receipt and return to testing program, and provide rules related to access. For CBT, procedures are in place to restrict access to the test delivery system by unauthorized personnel, track access by those with	External reviewers should only be able to access the pool of items to which they have been assigned. External parties involved in the development and/ or review of test items and forms should be required to sign confidentiality/ non-disclosure forms. Security of test materials at standard setting meetings and item review workshops may include the check in/out materials each day; show ID upon sign-in; restrict use of phones, laptops or other similar devices. *For assessments developed to be used in multiple states, if the is administered, scored and/or reported by a common vendor across states, many people may have access to the assessment materials at different points of the development, administration, scoring and reporting phases of the assessments. The testing program (e.g., SBAC, PARCC) is ultimately responsible for monitoring compliance with program-specified security standards to maintain the validity of the test score interpretations and uses. The quality of evidence related to this claim will rely on the extent to which the testing program can ensure the security of assessment materials.* *For assessments developed to be used in multiple states, if the assessment is not administered, scored and/or reported by a common vendor across states, this claim can only be partially evaluated using a common set of evidence (e.g., evidence related to security procedures during item development and review, and possibly assessment development). Claims related to the security of assessment materials during administration and scoring will need to be evaluated on a state-by state basis unless common guidelines/rules related to security procedures have been developed for use across all states regardless of vendor.*

	authorization, and maintain security in the log-in process. Procedures require the removal of materials from the testing location as soon as possible after administration, or, for CBT closing down of the test delivery system.	
E.1.4. Effective test security training is provided for all personnel who come into contact with test materials.	Training around security procedures and roles/ responsibilities related to maintaining test security is provided to all personnel at the testing program (i.e., employed directly by the test developer) and any authorized vendors. Test security training is provided to all test administrators, test coordinators, proctors and others who have access to secure test materials or test delivery systems. Test security training outlines the following: the purpose/need for test security procedures; specifications for handling secure materials and/or accessing secure, online testing sites; individual roles and responsibilities; and the consequences of non-compliance with test security procedures. Procedures are in place to ensure proctors/administrators receive appropriate training on how to detect, address, and report testing irregularities. A system is in place that allows for the identification of those who did/did not participate in test security training. Access is controlled so that those without training do not gain access to secure material.	Evidence should go beyond descriptions of security procedures and training specifications to reports or other documentation showing degree of participation by intended/required parties. For CBT, training includes details necessary to support secure, online administration such as disabling "screenshot" abilities and supervising test access through the use of secure student log-in identifiers. *For assessments developed to be used in multiple states, the trainings will need to be administered to each participating state and vendor.*
E.1.5. Procedures are in place to test and validate the effectiveness of security safeguards.	Procedures/analyses used to ensure the efficacy of different test security mechanisms are clearly articulated. Specifically, efficacy studies are in place for proctor training protocols, vendor chain of custody protocols, LEA chain of custody protocols, test administration manual descriptions of security (i.e., did readers understand the descriptions), detection of irregularities, and to ensure security of safeguards for computer-based and for paper-and-pencil tests, as relevant. The design and results of these procedures/ analyses are described in conjunction with how the information gained is used to inform practice.	For example, the utility of training/documentation related to test security is only realized if people take advantage of those resources. If direct evidence of training effectiveness is not provided, plan includes procedures to evaluate the effectiveness and clarity of test security training. These procedures may include: logging the number of calls/questions received related to test security during administration; collecting feedback regarding the clarity and usefulness of security training from participants; monitoring the number and type of testing irregularities that occur across sites. *For assessments developed to be used in multiple states, if the assessment is not administered, scored and/or reported by a common vendor some security procedures will be unique to each state and their selected vendors. In this case, the testing program may be able to provide evidence of procedures used to ensure the security of items during item*

		development and review, but the state would need to provide evidence that procedures put in place to ensure security in the administration of online and/ or paper-based form are effective.
E.1.6. Activities construed as cheating or other breaches of test security are clearly defined and transparent.	Test security requirements and training materials define the types of activities that will be considered cheating. The definition is broad enough to include previously unidentified methods by which test results could be artificially manipulated. Detailed procedures/specifications are in place regarding how such activities should be handled by proctors and the testing program during and after test administration. These definitions and procedures are easily accessible and publicly available (e.g., articulated in the test manual and/or online).	Definitions can account for the varying degree of severity of the cheating offence. As an example definition, the National Center for Education Statistics provides a breakdown of cheating offences in the following way: "Cheating in the first degree refers to willful and sometimes premeditated acts including: • Erasing and changing students' answers; • Filling in answers left blank by students; • Overtly and covertly providing correct answers on tests; • Falsifying student test identification or tracking numbers; and • Suspending or otherwise excluding students with poor academic performance on testing days, so that they are not tested. Cheating in the second degree includes more subtle forms of misconduct such as: • Cueing students on incorrect answers (for example, tapping on the desk or nudging); • Distributing 'cheat-sheets', talking students through processes and definitions; and • Giving extra time on tests during recess or before/after school" (USED, 2013, p. 3).[16] *For assessments developed to be used in multiple states, if states are using different vendors to support the administration, this claim will need to be evaluated on a state-by-state basis unless a common definition of cheating and how it is to be handled has been established for use across all states (regardless of vendor).*
E.1.7. Detailed procedures are in place to support the detection of testing irregularities.	Administration processes provide for the collection of data that can aid in detecting test irregularities (e.g., student test ID's should be linked to the name of the proctor, seating charts). Audit processes are in place to detect tampering before test administration and protocols are in place to support the reporting of expected/known security breaches (e.g., anonymous tip hotlines). Monitoring occurs to ensure items are not being compromised. Additionally, data forensics is employed to monitor item parameters and detect unusual drift in item characteristics and/or scoring protocols.	*For assessments developed to be used in multiple states, if states are using different vendors to support the administration, scoring and/or reporting, this claim will need to be evaluated on a state-by-state basis unless common guidelines/rules related to monitoring, detecting and reporting testing irregularities have been defined and are being used across all states (regardless of vendor).*

[16] U.S. Department of Education. (2013). Testing Integrity Symposium: Issues and Recommendations for Best Practice. Retrieved from: http://nces.ed.gov/pubs2013/2013454.pdf

	Multiple investigative and data analytic methods of detecting testing irregularities post-administration are in place to detect misconduct (e.g., erasure analyses, copying indices, gain score analyses). Data analysis plan specifies the flagging rules for each detection method and associated rationales. Legally defensible evidence is provided to support the selected methods for detecting testing irregularities and how they are used to make determinations regarding potential misconduct. There is no industry standard for flagging rules, but the rationale should reflect appropriate consideration of type-I and type-II errors. The procedures are consistent with the likelihood of issues occurring and the impact and consequences/costs of not detecting the cheating incident.	
E.1.8. Clearly documented procedures and specifications are provided for responding to breaches in test security.	In the case of exposed items, procedures provide clear guidance on response actions such as identifying the source of the breach, intervening if the item has been posted on the web, and analyses that will indicate the impact of the breach on scoring. If a full form has been exposed, breach or alternate forms using the same test blueprint and psychometric requirements are readily available. Root cause analysis and associated investigative procedures for following-up on identified irregularities are clear and fair. Procedures prioritize focus on the cases that are likely to be most severely in violation of the test security guidelines.	*For assessments developed to be used in multiple states, if states are using different vendors to support the administration, scoring and/or reporting, this claim will need to be evaluated on a state-by-state basis unless common procedures for responding to breaches in test security have been defined and are being used across all states (regardless of vendor).*

CCSSO Criterion A.2

The evaluation of evidence associated with A.2 involves judging the degree to which the assessment design and validity evaluation support the intended use and interpretation of assessment results. The primary claims related to this criterion are divided into two main sections:

1) Assessment Design: claims A.2.1-A.2.4 focus on the coherence between the assessment design and the stated purposes and uses of assessment results.
2) Validity Evaluation: claims A.2.5-A.2.7 focus on the quality and coherence of proposed/conducted validity studies and their results given the assumptions and inferences underlying the assessment design.

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

Relevant standards from the *Standards for Educational and Psychological Tests* (2014): 1.1, 1.2, 1.11, 12.4, 4.0, 4.1, 4.2, 1.6, 1.8, 1.9, 1.13, 1.16, 1.17, 1.18, 1.25, 12.2, 12.11, 13.3

Primary claims related to assessment design	Quality of Evidence	
	Sufficiency Statements	**Comments**
While no secondary claims are included directly, 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 the body of submitted evidence provides for all other criteria.		
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.	The overarching purpose(s) of the assessment, as a stand-alone measure, and within the context of a larger assessment or accountability system (where applicable) are clearly articulated. The population for which the test is intended is clearly defined including all demographic and experiential characteristics relevant to the test score interpretation and use. The specific set of uses and interpretations the assessment was designed to support is articulated. The intended use(s) are supported by the stated purpose(s) for assessment. The statements (or claims) to be made about students in light of assessment results (e.g., college and career readiness, mastery, proficiency, growth) in order to inform the intended uses are clearly defined.	*For assessments developed to be used in multiple states, the interpretations and uses the assessment was designed to support should be clearly articulated so it is clear to all states what does/does not fall within this category.*
A.2.2. The construct or content domain of interest, how it is defined, and the rationale for that specification are clearly articulated.	There is clear documentation of what the assessment is designed to measure. The documentation summarizes the scope of construct/extent of the content domain (i.e., knowledge and skills) to be assessed and how it was determined in light of the purpose of assessment and the manner in which results are intended to be used. The conceptual/empirical basis for the construct definition is described (e.g., reference to a theory, a set of standards, or some systematic analysis of the construct, domain or criterion given).	For example, the construct may be defined in a manner that prioritizes depth and a focus on the major shifts in the standards rather than coverage of the entire breadth of the standards. Similarly, if an assessment will be used to make inferences about student growth, the

	If multiple constructs/domains are intended to be measured, the expected relationship among them is described both qualitatively and empirically. Research studies are cited or proposed to document these relationships (e.g., phenomenological network, conceptual framework).	construct underlying those inferences, the different times at which measurement will occur, and the limitations of the scale (or assumptions about the scale, and logical/ empirical analysis of those assumptions, needed) for describing growth should be clear.
A.2.3. The assessment design reflects the construct definition and supports the intended interpretations and uses.	Documentation of a principled assessment design process is provided that highlights the connections among the design of the assessment (e.g., number, type, and cognitive demand of items, response format, number of scoring dimensions and their relationship to the content standards, scoring, etc.), the construct definition and the manner in which results are intended to be used. Key features of this documentation may include the following: • Evidence that the assessment design appropriately represents the construct definition - Demonstrated alignment between test blueprints/ assessment frameworks and the statements/claims the assessment is intended to support. - Rationales for the number and type of items/tasks used given the manner in which results are to be interpreted and used. - Feedback from experts that test blueprints and specifications are appropriate given the construct definition and intended goals/uses of results. • The population for which the test is intended was considered in the test design process (i.e., with respect to mode of administration, response requirements, section lengths, etc.). • Evidence that content and technical experts were integrally involved in the assessment design process.	Performance tasks may be necessary to make inferences about student demonstrations of deeper levels of cognition (e.g., Webb DOK levels 3 and 4). *For assessments developed to be used in multiple states, evidence should also include how diversity in the population of students to be tested across states was addressed in the assessment design.*
A.2.4. Documentation is provided that clearly specifies the inferences and assumptions underlying the design of the assessment.	A comprehensive theory of action, interpretive argument and/or other form of documentation is provided which summarizes: • The range of inferences and assumptions which must hold in order for assessment results to be interpreted and used in the manner intended. • The assumptions that must hold in order for the assessment development and implementation process, as well as the resulting test scores, to have the desired impact on systems and stakeholders. • This documentation may take a variety of different formats, but must include inferences associated with the accuracy and appropriateness of scoring procedures, the reliability and generalizability of results, the appropriateness of the measures for making specific types of decisions and other assumptions underlying the interpretation, use and access of test results in the manner intended.	

Primary claims related to validity evaluation	Quality of Evidence	
	Sufficiency Statements	**Comments**
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.[17]	Documentation is provided that details any evidence that was, or will be, collected to support each intended use and address the assumptions underlying the interpretive argument or validity evaluation plan for that use. The plan includes the proposed timeline for implementation. Specifications of the methodology used or planned to collect/ evaluate evidence provide sufficient detail to permit insight into the studies. Multiple areas of inquiry utilizing multiple lines of evidence are provided (e.g., test content, response processes, internal structure, relations to other variables, and consequences of testing). The inferences supported and the limitations of each study are clearly noted. A clear rationale is provided for the specific evidence collected to support each given inference/assumption (i.e., it clearly articulates why the evidence is important/relevant to support the use or defend underlying assumptions) in addition to any assumptions for which other relevant evidence was not/will not be obtained. The parties responsible for collecting evidence to support particular inferences/assumptions are documented. Summaries of research studies and their associated results include the composition of any sample of test takers from which validity evidence is obtained including major, relevant demographic and educational characteristics. This will indicate the degree of representativeness to the intended (sub) population of examinees. The research agenda and results are accessible and clear to allow stakeholders to make sound judgments about the quality of the proposed assessment system. This type of documentation will communicate information to multiple stakeholder groups including practitioners, researchers, educators and policymakers.	The testing program will only be responsible for those components necessary to support the design of the assessment and the uses it was *developed* to support. Evidence provided should include both descriptions of planned studies, and documentation /results from completed studies. *For assessments developed to be used in multiple states, the studies (including sampling plans) should address and account for the different policy and population contexts of each of the states administering the tests.*
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).	Rigorous studies (planned or completed) use representative samples and appropriate methodology to show that the content and response processes of the assessments appropriately represent the college and career readiness standards, including the cognitive demand of the standards (e.g., test blueprints demonstrate the learning progressions reflected in the content standards within and across grade levels, and experts in the content and progression toward readiness are significantly involved in the development process). The specified construct definition is supported with evidence. • Sub-score inter-correlations reflect the patterns expected given the type/manner of information they are intended to provide (unidimensional/multidimensional).	The type and manner of evidence provided to support this claim will vary depending on an assessment's phase of development or implementation. It is important to consider what type of evidence would be expected prior to administration vs. after the first operational exam vs. after 3 years of administration.

[17] It is understood that an assessment may not have been developed to support the uses specified in A.2.5.a- A.2.5.c. In this case, evidence supporting this use may not be provided, or may be limited. If a user intends to use a pre-existing assessment for a purpose it was not designed to support, it is their responsibility to obtain evidence to support that use. On the other hand, if a pre-existing assessment is being proposed by a vendor as appropriate for use in supporting a particular set of inferences, the onus to provide validity evidence falls to the developer.

		• Test results correlate with other assessments intended to measure a similar or related construct. • Test results are not related to other known measures considered unrelated or tangential to the assessment. For elementary and middle school assessments, studies are planned or conducted to show the relationship between performance at one grade and readiness to succeed at the next, which leads to success in meeting high school college and career readiness standards. For high school assessments, studies provide evidence related to the degree of relationship between student performance on the assessments and other agreed upon indicators of student success or readiness (e.g., readiness for credit-bearing courses at the start of college; probability of earning a C or better in related first year college courses). Rigorous studies with representative samples (qualifying the limitations where needed) are planned/have been implemented to provide evidence related to the degree of relationship between student performance on the assessments and other agreed upon indicators of student success or readiness. Evidence regarding the predictive validity of test performance for indicators of college and career readiness, or on track to CCR, is clear and substantial.	
	A.2.5.b. Evidence is provided to support the use of assessment results for making valid inferences about student growth over time.	The operational definition of growth (whether gains on a single construct or progression through a sequence of benchmarks) is clearly stated. Evidence is provided that the test was designed and scaled in consideration of the specified growth definition. A description and rationale is provided for any procedures used to estimate growth in light of student test scores (e.g., gain scores, SGPs, and VAMs). The assumptions on which the growth measures are based are explicit, as is evidence for those assumptions. Appropriate interpretations of these measures are clearly articulated. Studies include a methodology for establishing the degree of construct invariance across years in order to make appropriate (if qualified) inferences about student growth. Conducted or planned validation studies include evaluating growth measures in consideration of other indicators of student growth (e.g., grades, teacher perception surveys, interim/benchmark measures, etc.…). Observed relationships among the indicators of growth should be at least as strong as those reported in previous research unless qualified appropriately.	Growth should not be defined by performance on the assessment itself but rather be defined relative to the relevant knowledge, skills and abilities that students gain. *For assessments developed to be used in multiple states, the specific, common definition of growth the assessment was designed to support (for all states) should be clearly indicated.*

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 description and rationale is provided for any procedures used to estimate school, principal and/or teacher effectiveness estimates in light of assessment results (e.g., aggregate student growth metrics), and the manner in which they align with the purpose(s) of the test. The rationale includes an explanation of how sources of differences in score outcomes, other than school or teacher effectiveness, are eliminated or controlled for. Interpretations and limitations of these measures are specified. Validation studies include methodology for correlating effectiveness estimates based on test scores and other indicators of effectiveness (e.g., graduation rates, college enrollment rates, stakeholder surveys, course grades, credit hours earned). Results of validation studies reasonably reflect those relationships reported in existing research unless qualified appropriately. To the extent possible, evidence is provided that shows 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.2.6. The planned or completed validity evaluation considers the fairness of the assessment program for all examinees with respect to both intended and unintended consequences.	Planned or implemented validity studies provide evidence that assessments lead to the intended outcomes (i.e., meet the intended purposes specified in the theory of action) and minimize unintended negative consequences for all student sub-groups. Studies use representative samples and rigorous, appropriate methodology to provide evidence that intended score-based inferences are valid and fair for all individuals, especially those from student subgroups (e.g., students with disabilities, English learners). Documentation of sampling techniques and rationale are provided. Studies identify and provide recommendations for mitigating negative, unintended consequences of the test scores on instruction, student achievement, and other subsequent student outcomes including those consequences associated with both intended and unintended interpretations and uses.	The type of evidence expected will vary depending on the assessments phase of development or implementation. Prior to implementation, plans for validation and/or research conducted to ensure assessment items are fair and provide for desired inferences may be provided.
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).	Documentation is provided indicating the involvement of technical experts in the: • Design, review, and endorsement/approval of planned or implemented validity studies. • Review and endorsement of evidence collected to support the intended use and interpretation of assessment results. The relevant qualifications and experience of the experts is presented. Any deviations in the planned validity studies that have not been endorsed by experts are accompanied with rationales and evidence that the quality of the studies and their results was not sacrificed.	The quality of the documentation presented will depend on the extent to which the experts are independent, and the amount of evidence reviewed. External reviewers are preferable, to internal reviewers. However, internal, independent reviewers are preferable over less transparent quality control procedures.

REFERENCES

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

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.

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

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.