Annotated Assessment Peer Review Guidance

November 20, 2015

Prepared by

The Center for Assessment

For the

Council of Chief State School Officers

Based on the *Guidance* released by the U. S. Department of Education on September 25, 2015

Understanding this Document

In the *Peer Review of State Assessment Systems Non-Regulatory Guidance* (hereafter: *Guidance*) provided by the United States Department of Education (ED), there is a table for each critical element with two columns: *Critical Element* and *Examples of Evidence*. Our reading of the *Guidance* is that the *Examples of Evidence* column elaborates on what ED believes should be prioritized by peer reviewers in evaluating a state's submission. In this document, we replicate these two columns provided by ED and append additional columns and rows to help states interpret the text and to organize their peer review submissions.

We break down the *Critical Element* and *Examples of Evidence* columns into *Bite-sized Statements* states may consider using to support their peer review submissions in the *Bite-sized Statements* column. We also present a *Potential Sources* column to list sources from which evidence may be drawn to address the *Bite-sized Statements*. Finally, we present a *Notes* row to elaborate on dense or vague text presented in the *Critical Elements* and *Examples of Evidence* columns and on the types of evidence that may be found in the documents listed in the *Potential Sources* column. Bolded, superscript, capital letters in red (e.g., [A], [B], [C]) are used to denote where elaboration is provided in the *Notes* row.

This document is intended to serve as a resource to support states in preparing submissions. Other than text directly quoted from the *Guidance*, none of the information in this document should be regarded as official or final. The additional material reflects the judgments and interpretations of technical experts (in many cases informed by prior experienced with peer review). It should inform, but not, replace information provided directly from ED.

Notes:

- The *Bite-sized Statements* column represents our best interpretation of the components of an overall acceptable submission for a critical element based on the contents of the *Critical Element* and *Examples of Evidence* columns. We believe that our interpretations are defensible, but they do not supplant the judgment of ED and/or the peer reviewers.
- As noted in the *Guidance* (p. 15), the word "collectively" signals that peer review requires a summary and, potentially, a synthesis of the evidence for the critical element. That is, the evidence required by the element alone is insufficient. We have called attention to the word collectively where it appears (i.e., critical elements 2.5, 3.1, 4.1 and 6.4).
- Attention should be given to the use of phrases such as "evidence includes" versus "evidence may include" or "evidence such as." The latter cases indicate attempts to provide examples to promote clarity where multiple approaches may suffice.
- AA-AAS indicates Alternate Assessment-Alternate Achievement Standards.

Table of Contents

Critical Element 1.1 – State adoption of academic content standards for all student

Critical Element	Examples of Evidence	Bite-sized Statements

The State formally adopted challenging academic content standards for all students in reading/language arts, mathematics and science[A] and applies its academic content standards to all public elementary and secondary schools and students in the State[B].	Evidence to support this critical element for the State's assessment system includes:
	• Evidence of adoption of the State's academic content standards, specifically:
	  ○ Indication of *Requirement Previously Met*; or
	  ○ State Board of Education minutes, memo announcing formal approval from the Chief State School Officer to districts, legislation, regulations, or other binding approval of a particular set of academic content standards;
	• Documentation, such as text prefacing the State's academic content standards, policy memos, State newsletters to districts, or other key documents, that explicitly state that the State's academic content standards apply to all public elementary and secondary schools and all public elementary and secondary school students in the State[B];
	
	Note: A State with *Requirement Previously Met* should note the applicable category in the State Assessment Peer Review Submission Index for its peer submission. *Requirement Previously Met* applies to a State in the following categories: (1) a State that has academic content standards in reading/language arts, mathematics, or science that have not changed significantly since the State's previous assessment peer review; or (2)[C] with respect to academic content standards in reading/language arts and mathematics, a State approved for ESEA flexibility that (a) has adopted a set of college- and career-ready academic content standards that are common to a significant number of States and has not adopted supplemental State-specific academic content standards in these content areas, or (b) has adopted a set of college- and career-ready academic content standards certified by a State network of institutions of higher education (IHEs).
		1. Universal application of content standards[B]
		a. Standards cover all primary and secondary grade levels.
		b. Apply to all public elementary and secondary *schools*.
		c. Apply to all public elementary and secondary *students*.
		
		AND
		
		2A. Already approved content standards
		a. Standards have previously met peer review requirements and have not changed significantly since that time.
		
		OR
		
		2B. Recently adopted content standards[D]
		a. Adoption is legally binding.
		b. Standards cover reading/language arts, mathematics, and science.
		c. Standards are challenging[E].

Potential Sources of Evidence to Support Bite-sized Statements

1. Content standards, policy memos, newsletters, binding legislation or policy requiring compliance with ESEA.
2A. Peer review approval letter, ESEA flexibility approval letter.
2B. State Board of Education minutes, policy memos, content standards, standards development process.

Notes

A. Different documentation may be needed for each content area, depending upon whether the standards in each content area have changed since peer review and/or ESEA flexibility approval. For example, a state may have updated its reading/language arts and mathematics standards since receiving approval but not its science standards. In this case, a peer review or ESEA flexibility approval letter from ED would appear to suffice for science, but new evidence would be required for ELA and mathematics.
B. It appears from the ED wording that documentation that the standards apply to all elementary and secondary schools and students is separately required and that peer review or ESEA flexibility approval letter will not be sufficient. In addition to listed examples, this could take the form of legislation or policy requiring compliance with ESEA and/or documentation that school accountability for all public schools is based on assessments of the standards given to all students. ***Where districts are allowed to adopt their own content standards***, documentation showing that these statements are accurate will be necessary to obtain from each district.
C. A State with ESEA flexibility approval who has also changed its content standards in reading/language arts or mathematics since receiving that approval will need to show that the new standards meet the requirements detailed in point (2) of the notes in the *Examples of Evidence* column to retain flexibility approval and meet the requirements of peer review.
D. In content areas where existing standards have already received approval, the peer review or ESEA flexibility approval letter appears to be sufficient evidence. ***Where districts are allowed to adopt their own content standards***, documentation showing that these statements are accurate will be necessary to obtain from each district.
E. Procedures and/or policies put in place to assure that the content standards were challenging should be called out in the documentation of the standard development process.

Critical Element 1.2 – Coherent & rigorous academic content standards

Critical Element	Examples of Evidence	Bite-sized Statements
The State's academic content standards in reading/language arts, mathematics and science[A] specify what students are expected to know and be able to do by the time they graduate from high school to succeed in college and the workforce; contain content that is coherent (e.g., within and across grades) and rigorous; encourage the teaching of advanced skills; and were developed with broad stakeholder involvement.	Evidence to support this critical element for the State's assessment system includes: • Indication of *Requirement Previously Met;* or • Evidence that the State's academic content standards: ○ Contain coherent and rigorous content and encourage the teaching of advanced skills, such as: ▪ A detailed description of the **s**trategies the State used to ensure that its academic content standards adequately specify what students should know and be able to do; ▪ Documentation of the process used by the State to benchmark its academic content standards to nationally or internationally recognized academic content standards; ▪ Reports of external independent reviews of the State's academic content standards by content experts, summaries of reviews by educators in the State, or other documentation to confirm that the State's academic content standards adequately specify what students should know and be able to do; ▪ Endorsements or certifications by the State's network of institutions of higher education (IHEs), professional associations and/or the business community that the State's academic content standards represent the knowledge and skills in the content area(s) under review necessary for students to succeed in college and the workforce; ○ Were developed with broad stakeholder involvement, such as: ▪ Summary report of substantive involvement and input of educators, such as committees of curriculum, instruction, and content specialists, teachers and others, in the development of the State's academic content standards; ▪ Documentation of substantial involvement of subject-matter experts, including teachers, in the development of the State's academic content standards; ▪ Descriptions that demonstrate a broad range of stakeholders was involved in the development of the State's academic content standards, including individuals representing groups such as students with disabilities, English learners and other student populations in the State; parents; and the business community; ▪ Documentation of public hearings, public comment periods, public review, or other activities that show broad stakeholder involvement in the development or adoption of the State's academic content standards. Note: See note in Critical Element 1.1 – State Adoption of Academic Content Standards for All Students. With respect to academic content standards in reading/language arts and mathematics, *Requirement Previously Met* does not apply to supplemental State-specific academic content standards for a State approved for ESEA flexibility that has adopted a set of college- and career-ready academic content standards in a content area that are common to a significant number of States and also adopted supplemental State-specific academic content standards in that content area[B].	**1. State previously showed that its Standards are coherent and rigorous and developed with stakeholder involvement** a. State's standards have previously met peer review requirements and have not changed significantly since that time. ***OR both of the following*** **2A. Standards are Coherent and Rigorous** a. A defined and well documented process was used to develop the content standards. b. The standards are vertically articulated (coherent across grades)[C]. c. The standards are horizontally articulated (coherent within grades). d. The academic content standards adequately specify what students should know and be able to do. e. The academic content standards for high school adequately specify what students should know and be able to do to succeed in postsecondary endeavors such as college, career training, or the workforce. f. The standards are rigorous[D] and encourage teaching of advanced skills at all grade levels[E]. **2B. Standards were developed with stakeholder involvement** a. A broad range of stakeholders were involved in standards development process.

Critical Element 1.2 – Coherent & rigorous academic content standards

Potential Sources of Evidence to Support Bite-sized Statements
1. Peer review approval letter, ESEA flexibility approval letter 2. Standards development process documentation, reports of independent external review, summary of public comment, summary of revisions based on feedback, public notices requesting stakeholder involvement, benchmarking studies, independent external endorsements. 3. Standards development process documentation, public notices requesting stakeholder involvement, documentation of participant characteristics.

Notes
A. Different documentation may be needed for each content area depending upon whether the standards in each content area have changed since peer review and/or ESEA flexibility approval. For example, a state may have updated its reading/language arts and mathematics standards since receiving approval, but not its science standards. In this case, a peer review or ESEA flexibility approval letter would suffice for science, but new evidence would be required for ELA and mathematics. B. If a common set of standards was supplemented and/or modified, additional evidence will be needed. C. Vertically articulated (or vertically aligned) content standards are standards that have clear relationships across grades rooted in the ways student achievement becomes more advanced. That is, the content standards are related (or aligned) to each other across grades. For example, it could be shown that learning trajectories were considered in the standards development process, or that a formal process for evaluating changes across grades was incorporated. D. For example, benchmarking studies against standards nationally or internationally recognized as rigorous. For states that adopted standards substantially similar to other states, high-quality evidence for those standards provided by other states and an analysis of any differences is likely to be sufficient. E. For example, summary distributions of cognitive complexity levels in the standards show considerable proportions of standards elicit higher thinking levels and/or comparisons of distributions to previous standards showing an upward shift in the distributions of cognitive complexity levels in a few grade/subject combinations.

Critical Element 1.3 – Required assessments

Critical Element	Examples of Evidence	Bite-sized Statements
The State's assessment system includes annual general and alternate assessments (based on grade-level academic achievement standards or alternate academic achievement standards) in: • Reading/language arts and mathematics in each of grades 3-8 and at least once in high school (grades 10-12); • Science at least once in each of three grade spans (3-5, 6-9 and 10-12).	Evidence to support this critical element for the State's assessment system includes: • A list of the annual assessments the State administers in reading/language arts, mathematics and science including, as applicable, alternate assessments based on grade-level academic achievement standards or alternate academic achievement standards for students with the most significant cognitive disabilities, and native language assessments, and the grades in which each type of assessment is administered.	Not needed.

Potential Sources of Evidence
List or table of all relevant state assessments; test administration manuals; test coordinator manuals; binding statute, rule, or policy documents describing available and required assessments; binding statute, rule, or policy documents requiring compliance with ESEA.

Notes
Not needed.

Critical Element 1.4 – Policies for including all students in assessments

Critical Element	Examples of Evidence	Bite-sized Statements
The State requires the inclusion of all public elementary and secondary school students[A] in its assessment system and clearly and consistently communicates this requirement to districts and schools. • For students with disabilities, policies state that all students with disabilities in the State, including students with disabilities publicly placed in private schools as a means of providing special education and related services, must be included in the assessment system[A]; • For English learners: ○ Policies state that all English learners must be included in the assessment system, unless the State exempts a student who has attended schools in the U.S. for less than 12 months from one administration of its reading/ language arts assessment[B]; ○ If the State administers native language assessments, the State requires English learners to be assessed in reading/language arts in English if they have been enrolled in U.S. schools for three or more consecutive years, except if a district determines, on a case-by-case basis[C], that native language assessments would yield more accurate and reliable information, the district may assess a student with native language assessments for a period not to exceed two additional consecutive years[D].	Evidence to support this critical element for the State's assessment system includes documents such as: • Key documents, such as regulations, policies, procedures, test coordinator manuals, test administrator manuals and accommodations manuals that the State disseminates to educators (districts, schools and teachers), that clearly state that all students must be included in the State's assessment system and do not exclude any student group or subset of a student group[A]; • For students with disabilities, if needed to supplement the above: Instructions for Individualized Education Program (IEP) teams and/or other key documents; • For English learners, if applicable and needed to supplement the above: Test administrator manuals and/or other key documents that show that the State provides a native language (e.g., Spanish, Vietnamese) version of its assessments.	**1. All students are included in the assessment system.**[A] a. All public elementary and secondary school students are required to participate in the assessment system. b. All students with disabilities, including those publically placed into private schools, are required to participate in the assessment system. c. All English Language learners are required to participate in the assessment system with exemption policies clearly explicated. **2. EL exceptions** a. EL exemptions from testing apply to only reading/language arts, and only for students that have attended U.S. schools for less than 12 months. b. Case-by-case extensions of the three-year limit on native language testing are capped at two years[C]. **3. Communications** a. The requirement that all public elementary and secondary school students are included in the State assessment system is clearly communicated to districts and schools. b. Inclusion policies for SWDs and ELs are clearly communicated to districts and schools

Potential Sources of Evidence to Support Bite-sized Statements

1 & 2: Binding statute, regulation, rule, or policy; binding requirement to comply with ESEA; test coordinator manuals; accommodations manuals; test training materials; technical manuals; requirements, training materials, and/or templates for IEP/504 plan/EL plan teams.
3: Official communications to districts and schools

Notes

A. For states where parental refusal (opt-out) is allowable, states may need to explain attempts to balance the preservation of inclusion policies and parental rights. ED has indicated that the 95% participation requirement remains and that two avenues of action may be taken: (1) review by peers of the adequacy of state guidance on participation, and (2) review of state compliance outside of peer review. However, ED has also indicated that the effect of having less than 95% participation on peer review is not yet clear.
B. We assume documentation will be needed to show compliance. Examples of this documentation include a section of a technical report that analyzes exemption rates for reading/language arts or documentation of policies or other state mechanisms that enforce inclusion.
C. It is unclear whether ED expects states to document state oversight of the case-by-case decisions of districts.
D. We assume documentation of compliance with the two-year extension maximum will be needed. For example, a section of a technical report that summarizes the number of consecutive years students have taken a native language assessment or documentation of policies or other state mechanisms that enforce inclusion.

Critical Element 1.5 Participation data

<table>
<tr><th>Critical Element</th><th>Examples of Evidence</th><th>Bite-sized Statements</th></tr>
<tr>
<td>The State's participation data show that all students, disaggregated by student group and assessment type, are included in the State's assessment system. In addition, if the State administers end-of-course assessments for high school students, the State has procedures in place for ensuring that each student is tested and counted in the calculation of participation rates on each required assessment and provides the corresponding data.</td>
<td>
Evidence to support this critical element for the State's assessment system includes:

Participation data from the most recent year of test administration in the State, such as in Table 1 below, that show that all students, disaggregated by student group (i.e., students with disabilities, English learners, economically disadvantaged students, students in major racial/ethnic categories, migratory students, and male/female students) and assessment type (i.e., general and AA-AAS) in the tested grades are included in the State's assessments for reading/language arts, mathematics and science[A];
If the State administers end-of-course assessments for high school students, evidence that the State has procedures in place for ensuring that each student is included in the assessment system during high school, including students with the most significant cognitive disabilities who take an alternate assessment based on alternate academic achievement standards and recently arrived English learners who take an ELP assessment in lieu of a reading/language arts assessment[B], such as:

Description of the method used for ensuring that each student is tested and counted in the calculation of participation rate on each required assessment. If course enrollment or another proxy is used to count all students, a description of the method used to ensure that all students are counted in the proxy measure;
Data that reflect implementation of participation rate calculations that ensure that each student is tested and counted for each required assessment. Also, if course enrollment or another proxy is used to count all students, data that document that all students are counted in the proxy measure.

Table 1: Students Tested by Student Group in [subject] during [school year]
<table>
<tr><th>Student group</th><th></th><th>Gr 3</th><th>Gr 4</th><th>Gr 5</th><th>Gr 6</th><th>Gr 7</th><th>Gr 8</th><th>HS</th></tr>
<tr><td rowspan="3">All</td><td># enrolled</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td># tested</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>% tested</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td rowspan="3">Economically disadvantaged</td><td># enrolled</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td># tested</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>% tested</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td rowspan="3">Students with Disabilities</td><td># enrolled</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td># tested</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>% tested</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td rowspan="3">(Continued for for all others student groups)</td><td># enrolled</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td># tested</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>% tested</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
</table>
Number of students assessed on the State's AA-AAS in [subject] during [school year]: _____.

Note: A student with a disability should only be counted as tested if the student received a valid score on the State's general or alternate assessments submitted for assessment peer review for the grade in which the student was enrolled. If the State permits a recently arrived English learner (i.e., an English learner who has attended schools in the U.S. for less than 12 months) to be exempt from one administration of the State's reading/language arts assessment and to take the State's ELP assessment in lieu of the State's reading/language arts assessment, then the State should count such students as 'tested' in data submitted to address critical element.
</td>
<td>
1. Participation

a. The most recent state assessment participation data (disaggregated by grade, assessment type and accountability student groups) show that all students are included in the state assessment system[C].
b. States with end-of-course assessment configurations for high school: each student's participation is appropriately captured in high-school participation[D] metrics.

</td>
</tr>
</table>

Critical Element 1.5 Participation data

Potential Sources of Evidence to Support Bite-sized Statements

Technical reports, business rules for calculating assessment participation, public assessment reports, binding policies for inclusion in end of course high school assessments.

Notes

A. Table 1 does not appear to incorporate assessment type (e.g., general vs. AA-AAS). It may be necessary to revise or create different versions of the table to show participation by assessment type. For example, for the "All" Student group in Table 1 could be modified as:

Student Group	Statistic	General Assessment							Alternate Assessment							All Assessments						
		Gr 3	Gr 4	Gr 5	Gr 6	Gr 7	Gr 8	HS	Gr 3	Gr 4	Gr 5	Gr 6	Gr 7	Gr 8	HS	Gr 3	Gr 4	Gr 5	Gr 6	Gr 7	Gr 8	HS
All	# enrolled																					
	# tested																					
	% tested																					
Economically disadvantaged	# enrolled																					
	# tested																					
	% tested																					
Students with Disabilities	# enrolled																					
	# tested																					
	% tested																					
(Continued…)	# enrolled																					
	# tested																					
	% tested																					

B. This appears to be intended to assure that participation calculations capture every student at some point during high school for each required test.

C. For states where parental refusal (opt-out) is allowable, states may need to explain attempts to balance the preservation of inclusion policies and parental rights. ED has indicated that the 95% participation requirement remains, and that two avenues of action may be taken: (1) review by peers of the adequacy of state guidance on participation, and (2) review of state compliance outside of peer review. However, ED has also indicated that the effect of having less than 95% participation on peer review is not yet clear

D. As explained in the note, English Learners who are permitted to take an ELP assessment in lieu of the State's reading/language arts assessment should be counted as tested.

Critical Element 2.1 – Test Design and Development[A]

Critical Element	Examples of Evidence	Bite-sized Statements

The State's test design and test development process is well-suited for the content, is technically sound, aligns the assessments to the full range[B,C] of the State's academic content standards, and includes:
• Statement(s) of the purposes of the assessments and the intended interpretations and uses of results;
• Test blueprints that describe the structure of each assessment in sufficient detail to support the development of assessments that are technically sound, measure the full range of the State's grade-level academic content standards, and support the intended interpretations and uses of the results;
• Processes to ensure that each assessment is tailored to the knowledge and skills included in the State's academic content standards, reflects appropriate inclusion of challenging content, and requires complex demonstrations or applications of knowledge and skills (i.e., higher-order thinking skills);
• If the State administers computer-adaptive assessments, the item pool and item selection procedures adequately support the test design.
	Evidence to support this critical element for the State's general assessments and AA-AAS includes:
	
	For the State's general assessments:
	• Relevant sections of State code or regulations, language from contract(s) for the State's assessments, test coordinator or test administrator manuals, or other relevant documentation that states the purposes of the assessments and the intended interpretations and uses of results;
	• Test blueprints that:
	  ○ Describe the structure of each assessment in sufficient detail to support the development of a technically sound assessment, for example, in terms of the number of items, item types, the proportion of item types, response formats, range of item difficulties, types of scoring procedures, and applicable time limits;
	  ○ Align to the State's grade-level academic content standards in terms of content (i.e. knowledge and cognitive process), the full range of the State's grade-level academic content standards, balance of content, and cognitive complexity;
	• Documentation that the test design that is tailored to the specific knowledge and skills in the State's academic content standards (e.g., includes extended response items that require demonstration of writing skills if the State's reading/language arts academic content standards include writing)[B,C];
	• Documentation of the approaches the State uses to include challenging content and complex demonstrations or applications of knowledge and skills (i.e., items that assess higher-order thinking skills, such as item types appropriate to the content that require synthesizing and evaluating information and analytical text-based writing or multiple steps and student explanations of their work); for example, this could include test specifications or test blueprints that require a certain portion of the total score be based on item types that require complex demonstrations or applications of knowledge and skills and the rationale for that design.
	
	  For the State's technology-based general assessments, in addition to the above:
	• Evidence of the usability of the technology-based presentation of the assessments, including the usability of accessibility tools and features (e.g., embedded in test items or available as an accompaniment to the items), such as descriptions of conformance with established accessibility standards and best practices and usability studies;
	• For computer-adaptive general assessments:
	  ○ Evidence regarding the item pool, including:
	  ▪ Evidence regarding the size of the item pool and the characteristics (non-statistical (e.g., content) and statistical) of the items it contains that demonstrates that the item pool has the capacity to produce test forms that adequately reflect the State's test blueprints in terms of:
	  - Full range of the State's academic content standards, balance of content, cognitive complexity for each academic content standard, and range of item difficulty levels for each academic content standard;
	  - Structure of the assessment (e.g., numbers of items, proportion of item types and response types);
		1. Test Purposes
		a. Each intended interpretation and use of assessment results is clearly articulated[C].
		2. Assessment Structure
		a. Test blueprints are aligned to the state content standards[C,E] in terms of both the knowledge[F] and cognitive complexity[G], with respect to both overall coverage[H] and balance of coverage.[I]
		b. All test forms are reflective of the test blueprint, and are thus aligned to the state content standards.[J]
		c. The distribution of cognitive complexity for each test form and content area is sufficient to assess the depth and complexity of the state's standards (i.e., the assessment measures higher-order cognitive processes).[K]
		d. The distribution of item types for each test form and content area is sufficient to assess the depth and complexity of the standards being addressed.[L]
		e. **Computer Adaptive Assessments:** The item pool and item selection procedures assure that each student receives a test that (i) assesses the full depth and breadth of the state content standards[M], and (ii) allows for adequately precise estimates of student achievement.
		f. **AA-AAS:** There is a clear and documented link between the general standards and the alternate standards, with a sound rationale for reduction in depth,

Critical Element 2.1 – Test Design and Development[A]

Critical Element	Examples of Evidence	Bite-sized Statements
	○ Technical documentation for item selection procedures that includes descriptive evidence and empirical evidence (e.g., simulation results that reflect variables such as a wide range of student behaviors and abilities and test administration early and late in the testing window) that show that the item selection procedures are designed adequately for: ▪ Content considerations to ensure test forms that adequately reflect the State's academic content standards in terms of the full range of the State's grade-level academic content standards, balance of content, and the cognitive complexity for each standard tested; ▪ Structure of the assessment specified by the blueprints; ▪ Reliability considerations such that the test forms produce adequately precise estimates of student achievement for all students (e.g., for students with consistent and inconsistent testing behaviors, high- and low-achieving students; English learners and students with disabilities); ▪ Routing students appropriately to the next item or stage; ▪ Other operational considerations, including starting rules (i.e., selection of first item), stopping rules, and rules to limit item over-exposure. **AA-AAS.** For the State's AA-AAS[D]: • Relevant sections of State code or regulations, language from contract(s) for the State's assessments, test coordinator or test administrator manuals, or other relevant documentation that states the purposes of the assessments and the intended interpretations and uses of results for students tested; • Description of the structure of the assessment, for example, in terms of the number of items, item types, the proportion of item types, response formats, types of scoring procedures, and applicable time limits. For a portfolio assessment, the description should include the purpose and design of the portfolio, exemplars, artifacts, and scoring rubrics; • Test blueprints (or, where applicable, specifications for the design of portfolio assessments) that reflect content linked to the State's grade-level academic content standards and the intended breadth and cognitive complexity of the assessments; • To the extent the assessments are designed to cover a narrower range of content than the State's general assessments and differ in cognitive complexity: ○ Description of the breadth of the grade-level academic content standards the assessments are designed to measure, such as an evidence-based rationale for the reduced breadth within each grade and/or comparison of intended content compared to grade-level academic content standards; ○ Description of the strategies the State used to ensure that the cognitive complexity of the assessments is appropriately challenging for students with the most significant cognitive disabilities; ○ Description of how linkage to different content across grades/grade spans and vertical articulation of academic expectations for students is maintained; • If the State developed extended academic content standards to show the relationship between the State's grade-level academic content standards and the content of the assessments, documentation of their use in the design of the assessments; • For adaptive alternate assessments (both computer-delivered and human-delivered), evidence, such	breadth, and/or complexity; and the alternate content standards retain the maximum complexity appropriate for the AA-AAS population.[N] g. **AA-AAS:** The usability and appropriateness of the test mode and test items (including accessibility features) are adequate.[O]

Critical Element 2.1 – Test Design and Development[A]

Critical Element	Examples of Evidence	Bite-sized Statements
	as a technical report for the assessments, showing: ○ Evidence that the size of the item pool and the characteristics of the items it contains are appropriate for the test design; ○ Evidence that rules in place for routing students are designed to produce test forms that adequately reflect the blueprints and produce adequately precise estimates of student achievement for classifying students; ○ Evidence that the rules for routing students, including starting (e.g., selection of first item) and stopping rules, are appropriate and based on adequately precise estimates of student responses, and are not primarily based on the effects of a student's disability, including idiosyncratic knowledge patterns; ○ For technology-based AA-AAS, in addition to the above, evidence of the usability of the technology-based presentation of the assessments, including the usability of accessibility tools and features (e.g., embedded in test items or available as an accompaniment to the items), such as descriptions of conformance with established accessibility standards and best practices and usability studies.	

Potential Sources of Evidence to Support Bite-sized Statements

For both general and alternate assessments:

1. Legislation, policies, or other official documents stating assessment purposes; official communications to stakeholders regarding assessment purposes.
2. Test specifications, test blueprints, test development procedures, technical manuals, item specifications, alignment studies, CAT administration rules, CAT item selection procedures, CAT simulation procedures, CAT simulation rules, CAT item sufficiency studies, alternate content standards development process, process and rationale for determining the structure of the test (e.g., educator advisory committee process for how various sections of the content standards are weighted for importance).

Notes

A. This critical element appears to be a compilation and summary of evidence from other critical elements addressing test purposes (3.1), alignment of items to content standards and blueprint (2.2), alignment of test forms (4.5, 4.6), and CAT item pool and item selection (3.1, 4.1, 4.5, 4.6).
B. This likely requires use of more than multiple-choice items for standards describing high-level cognitive skills such as those describing high-level problem solving in mathematics, inquiry in science, and synthesis in all disciplines.
C. It is no longer acceptable to assess writing only once per grade span for states with writing content standards in each grade. Whether locally-administered assessments of writing could fulfill this requirement depends on whether those assessments can meet the requirements in this peer review guidance. ED has indicated that because of current technical difficulties, there is some flexibility for not assessing listening and speaking if they are included in a state's language arts standards. If speaking and/or listening are included in state content standards but not in state assessment, a waiver will need to submitted to ED. Details of the waiver have not been released, but it is known that the waiver will require a plan for speaking and listening to be incorporated in state assessment in the future.
D. Accessibility considerations should be woven throughout the item and test development process. Issues of accessibility should be considered in every step of the test development and administration process.
E. States administering common assessments (e.g., consortia tests, ACT, SAT), that have also created State-specific standards in assessed grades/courses to supplement the common standards, will need to provide evidence to show how that difference is attended to in State's assessments.
F. That is, the categories of content knowledge in the content standards.
G. That is, the levels of cognitive complexity in the standards (typically using a discipline-specific Depth of Knowledge indicator, but other taxonomies of levels of complexity in cognitive processes may be used).
H. That is, blueprints reflect coverage of the full range of the content standards (not necessarily that every content standard is represented by at least one item, but that the items

Critical Element 2.1 – Test Design and Development[A]

Notes
reflect the full range of the standards). I. That is, test blueprints are balanced in representing different parts of the content standards (e.g., there is no substantial over- or under-representation of subsets of the content standards). J. Could be based on studies of the alignment of the set of items on each of various forms to the intended blueprint or to the content standards themselves. It may not be necessary to study alignment of every form in every grade and subject, but to demonstrate that the processes for item and test form development leads to the administration of aligned tests. K. Although this is addressed implicitly as part of previous *Bite-sized Statements*, the wording of the guidance suggests that this specific aspect should be explicitly and separately addressed. Although this statement does not explicitly require performance tasks, it will likely be challenging to meet this requirement with an assessment that is entirely selected response. Item types that elicit more complex knowledge and skills can be a key component of the state's evidence. ED has indicated that the important evaluation is that the assessment measures the full range (depth and breadth) of the state's adopted standards. ***It is important to note that this does not require that all items measure higher-order thinking skills***. This applies only to items intended to measure content standards describing higher-order thinking skills. L. Although appropriate use of item types to achieve alignment is addressed implicitly in previous *Bite-sized Statements*, the wording of the guidance implies that this should be explicitly and separately addressed. M. Requires that all students experience a test consistent with the same blueprint, even though it may be that no student will receive the same test as any other. It is important to demonstrate that the item selection procedures prioritize meeting the blueprint and achieving alignment. N. Much of the evidence for this *Bite-sized Statement* can be drawn from evidence developed for critical elements 1.1 and 1.2. O. Strongly related to critical element 3.1.

Critical Element 2.2 – Item development

Critical Element	Examples of Evidence	Bite-sized Statements
The State uses reasonable and technically sound procedures to develop and select items to assess student achievement based on the State's academic content standards in terms of content and cognitive process, including higher-order thinking skills.	Evidence to support this critical element for the State's general assessments and AA-AAS includes documents such as: For the State's general assessments, evidence, such as a sections in the technical report for the assessments, that show: • A description of the process the State uses to ensure that the item types (e.g., multiple choice, constructed response, performance tasks, and technology-enhanced items) are tailored for assessing the academic content standards in terms of content; • A description of the process the State uses to ensure that items are tailored for assessing the academic content standards in terms of cognitive process (e.g., assessing complex demonstrations of knowledge and skills appropriate to the content, such as with item types that require synthesizing and evaluating information and analytical text-based writing or multiple steps and student explanations of their work); • Samples of item specifications that detail the content standards to be tested, item type, intended cognitive complexity, intended level of difficulty, accessibility tools and features, and response format; • Description or examples of instructions provided to item writers and reviewers; • Documentation that items are developed by individuals with content area expertise, experience as educators, and experience and expertise with students with disabilities, English learners, and other student populations in the State; • Documentation of procedures to review items for alignment to academic content standards, intended levels of cognitive complexity, intended levels of difficulty, construct-irrelevant variance, and consistency with item specifications, such as documentation of content and bias reviews by an external review committee; • Description of procedures to evaluate the quality of items and select items for operational use, including evidence of reviews of pilot and field test data; • As applicable, evidence that accessibility tools and features (e.g., embedded in test items or available as an accompaniment to the items) do not produce an inadvertent effect on the construct assessed; • Evidence that the items elicit the intended response processes, such as cognitive labs or interaction studies. **AA-AAS.** For the State's AA-AAS, in addition to the above: • If the State's AA-AAS is a portfolio assessment, samples of item specifications that include documentation of the requirements for student work and samples of exemplars for illustrating levels of student performance; • Documentation of the process the State uses to ensure that the assessment items are accessible, cognitively challenging, and reflect professional judgment of the highest achievement standards possible for students with the most significant cognitive disabilities.  For the State's technology-based general assessments and AA-AAS:	**1. Item Development** a. Technically sound[C] procedures and criteria[A] are clearly described to align items to intended content standards content and cognitive complexity[B]. b. Procedures described were followed, and criteria were used to determine when item development was complete[D]. c. Items were successfully developed to match intended content and cognitive complexity, including higher-order thinking[E]. d. Accessibility tools, where available, did not introduce construct-irrelevant variance. e. **AA-AAS**: Items are accessible. f. **AA-AAS**: Items are cognitively challenging and reflect the highest appropriate level of achievement for the AA-AAS population[F]. g. **AA-AAS**: for portfolio programs, educators are provided with sample item specifications and exemplars at each performance level. **2. Item selection** a. Technically sound procedures and criteria are clearly described to maximize alignment between the total set of items and the content standards/content specifications.[G] b. Technically sound procedures and criteria for appropriate item functioning[H] in pilot and/or field test data are clearly described. c. Procedures described were followed and criteria described were used in item selection. d. The total sets of items on test forms are strongly aligned to content standards/specifications. e. **AA-AAS:** Items were evaluated for appropriate accessibility and usability.[I] f. **Computer-Based Assessment**: Items were evaluated for appropriate display, accessibility, and usability, including use of accommodations, and electronic delivery via allowable devices.[I] g. **Computer-Based Assessment**:There is consistency of item display and usability across

Critical Element 2.2 – Item development

Critical Element	Examples of Evidence	Bite-sized Statements
	• Documentation that procedures to evaluate and select items considered the deliverability of the items (e.g., usability studies). Note: This critical element is closely related to Critical Element 4.2 – Fairness and Accessibility.	accommodations and devices. h. **CAT**: The item pool facilitates matching content and cognitive skills in the standards on test forms even for students at very low and high skill levels.[J]

Potential Sources of Evidence to Support Bite-sized Statements

1. Technical manuals, alignment studies, usability studies, content specifications, item/task specifications, test specifications, student work exemplars at various performance levels, training materials, instructions for item writers and reviewers, item writing and review procedure documents, item statistics summaries, invariance studies, cognitive lab reports. Audit trails showing application of procedures and criteria may also be helpful.
2. Technical manuals, alignment studies, test blueprints, tem selection specifications and procedures. Audit trails showing application of procedures and criteria may also be helpful.

Notes

A. For example, this could be supported from some combination of the following:
   - Procedures and criteria to determine what item type(s) to use to measure specific content standards/content specifications.
   - Training provided to (i) stimulus and item developers, (ii) content/process review committees, (iii) bias/sensitivity review committees, and/or (iv) accessibility/usability review committees.
   - Procedures used by stimulus and item developers to assert alignment before submission.
   - Procedures used by content review committees to evaluate match between items and content standards.
   - Procedures and criteria for recruiting, screening, and evaluating the work of (i) stimulus and item developers and/or (ii) item reviewers on each of the different item review committees with documentation of use and degree of success in achieving the criteria.
   - Procedures used to achieve representation of ELs, SWDs, and various demographic groups on stimulus and item development committees and item review committees (and documentation of degree of success)

B. Typically measured using Webb's Depth of Knowledge (DOK) ratings, but other taxonomies are used.

C. By technically sound, we mean that the procedures and criteria are considered best practice and result in item-level alignment to intended content standards and cognitive complexity. For example, it could be shown how procedures and criteria are consistent with the AERA/APA/NCME *standards*, CCSSO/ATP *operational best practices*, or other current research on best practices.

D. This could be demonstrated by a statement or documentation from the development vendor providing assurances that processes were followed.

E. For example, summary statistics showing the degree of match between items and content standards, both in terms of content and cognitive complexity.

F. For example, could be argued from
   - Training provided to stimulus and item developers
   - Training provided to item reviewers
   - Procedures used by stimulus and item developers to assert development to the highest appropriate achievement level.
   - Procedures used by item reviewers to evaluate whether items are written at the highest appropriate achievement level.

G. Both in terms of content covered and cognitive complexity.

H. E.g., that the item statistics (both classical and item response theory values) fall within acceptable ranges (i.e., within the specifications).

I. It should not be necessary to evaluate every item for usability, but a sample of items sufficiently representative of item types, formats, and other variables to support a claim for the entire set of selected items.

J. For example, charts showing pool information across the range of student achievement for all major content categories and levels of cognitive complexity, and/or explanation of an item selection algorithm that assures match to the content standards even if a highly informative item is not available.

Critical Element 2.3 – Test administration

Critical Element	Examples of Evidence	Bite-sized Statements
The State implements policies and procedures for standardized test administration, specifically the State: • Has established and communicates to educators clear, thorough and consistent standardized procedures for the administration of its assessments, including administration with accommodations; • Has established procedures to ensure that all individuals responsible for administering the State's general and alternate assessments receive training on the State's established procedures for the administration of its assessments; • If the State administers technology-based assessments, the State has defined technology and other related requirements, included technology-based	Evidence to support this critical element for the State's general assessments and AA-AAS includes: • Regarding test administration: ○ Test coordinator manuals, test administration manuals, accommodations manuals and/or other key documents that the State provides to districts, schools, and teachers that address standardized test administration and any accessibility tools and features available for the assessments; ○ Instructions for the use of accommodations allowed by the State that address each accommodation. For example: ▪ For accommodations such as bilingual dictionaries for English learners, instructions that indicate which types of bilingual dictionaries are and are not acceptable and how to acquire them for student use during the assessment; ▪ For accommodations such as readers and scribes for students with disabilities, documentation of expectations for training and test security regarding test administration with readers and scribes; ○ Evidence that the State provides key documents regarding test administration to district and school test coordinators and administrators, such as e-mails, websites, or listserv messages to inform relevant staff of the availability of documents for downloading or cover memos that accompany hard copies of the materials delivered to districts and schools; ○ Evidence of the State's process for documenting modifications or disruptions of standardized test administration procedures (e.g., unapproved non-standard accommodations, electric power failures or hardware failures during technology-based testing), such as sample of incidences documented during the most recent year of test administration in the State. • Regarding training for test administration: ○ Evidence regarding training, such as: ▪ Schedules for training sessions for different groups of individuals involved in test administration (e.g., district and school test coordinators, test administrators, school computer lab staff, accommodation providers); ▪ Training materials, such as agendas, slide presentations and school test coordinator manuals and test administrator manuals, provided to participants. For technology-based assessments, training materials that include resources such as practice tests and/or other supports to ensure that test coordinators, test administrators and others involved in test administration are prepared to administer the assessments; ▪ Documentation of the State's procedures to ensure that all test coordinators, test administrators, and other individuals involved in test administration receive training for each test administration, such as forms for sign-in sheets or screenshots of electronic forms for tracking attendance, assurance forms, or identification of individuals responsible for tracking attendance.  For the State's technology-based assessments: • Evidence that the State has clearly defined the technology (e.g., hardware, software, internet connectivity, and internet access) and other related requirements (e.g., computer lab configurations) necessary for schools to administer the assessments and has communicated these requirements to schools and districts; • District and school test coordinator manuals, test administrator manuals and/or other key documents that include specific instructions for administering technology-based assessments (e.g., regarding necessary advanced preparation, ensuring that test administrators and students are adequately familiar with the delivery devices and, as applicable, accessibility tools and features available for students);	1. **Standardization** a. Test distribution and administration directions are clear and consistent (including administration with accommodations) and clearly communicated to test administrators. b. Procedures are clear regarding the documentation of irregularities and are clearly communicated to districts and schools.[B] c. Procedures are established for requesting and receiving accommodations and are clearly communicated to educators and parents (e.g., IEP teams).[C] **2. Training** a. Test coordinators and administrators have had sufficient training for administering the assessments (including with accommodations) and for reporting irregularities or disruptions.[D] b. AA-AAS: Test administrators have been appropriately trained and have clear instructions for administering alternate assessments, including the administration of items that require their involvement. **3. Computer-based assessment** a. The technology

Critical Element 2.3 – Test administration

Critical Element	Examples of Evidence	Bite-sized Statements
test administration in its standardized procedures for test administration, and established contingency plans to address possible technology challenges during test administration.	• Contingency plans or summaries of contingency plans that outline strategies for managing possible challenges or disruptions during test administration. **AA-AAS.** For the State's AA-AAS, in addition to the above: • If the assessments involve teacher-administered performance tasks or portfolios, key documents, such as test administration manuals, that the State provides to districts, schools and teachers that include clear, precise descriptions of activities, standard prompts, exemplars and scoring rubrics, as applicable; and standard procedures for the administration of the assessments that address features such as determining entry points, selection and use of manipulatives, prompts, scaffolding, and recognizing and recording responses; • Evidence that training for test administrators addresses key assessment features, such as teacher-administered performance tasks or portfolios; determining entry points; selection and use of manipulatives; prompts; scaffolding; recognizing and recording responses; and/or other features for which specific instructions may be needed to ensure standardized administration of the assessment.	requirements are clearly specified and clearly communicated to schools and districts. b. Technology issues are clearly addressed in training and manuals.[E] c. In the case of test interruptions, sound contingency plans are available for dealing appropriately with interruptions.

Potential Sources of Evidence to Support Bite-sized Statements

1. Test coordinator manuals, test administration manuals, accommodations manuals, training materials, official communications to educators
2. Training materials, documentation of training attendance
3. Technology requirements and specifications, official communications to educators, test coordinator manuals, training materials

Notes

A. States may have multiple vendors for the multiple assessments administered. The State needs to show they have thought about a way to present information from different vendors (e.g., test administration instructions) in a way that is clear and consistent across the assessment system.
B. For example, unapproved accommodations or interruptions to testing.
C. These procedures should be communicated well in advance of testing so there is enough time to determine appropriate student-accommodations/accessibility features match at the school level.
D. ED has communicated that they would rather see the actual documents used for training rather than any narrative prepared for the purposes of peer review.
E. For example, preparing for and monitoring online assessment, assuring students are familiar with devices and the test platform, and references to documentation of technical requirements for districts and schools in order to minimize the possibility of test interruptions.

Critical Element 2.4 – Monitoring test administration [A]

Critical Element	Examples of Evidence	Bite-sized Statements
The State adequately monitors the administration of its State assessments to ensure that standardized test administration procedures are implemented with fidelity across districts and schools.	Evidence[B] to support this critical element for the State's general assessments and AA-AAS includes documents such as: • Brief description of the State's approach to monitoring test administration (e.g., monitoring conducted by State staff, through regional centers, by districts with support from the State, or another approach); • Existing written documentation of the State's procedures for monitoring test administration across the State, including, for example, strategies for selection of districts and schools for monitoring, cycle for reaching schools and districts across the State, schedule for monitoring, monitors' roles, and the responsibilities of key personnel; • Summary of the results of the State's monitoring of the most recent year of test administration in the State.	***1. Test Administration*** a. Effective procedures[C] are in place to monitor[D] test administration. b. The test administration monitoring procedures have been carried out with a high degree of fidelity.

Potential Sources of Evidence to Support Bite-sized Statements

Procedures for monitoring administration, training materials for monitors, records of administration monitoring, process and tools for self-reporting anomalies, help desk training materials, help-desk logs, summaries of monitoring reports, summaries of self-reports, summaries of monitoring appeals, communications to schools and districts regarding both physical and electronic monitoring.

Notes

A. This critical element is a "checklist" item and will not go to the peer reviewers unless ED finds an issue with the evidence submitted.
B. The wording of this critical element and examples of evidence suggest that at least some on-site monitoring (i.e., monitors physically present in testing environments) will be necessary.
C. If the state monitors districts/schools, procedures should detail (1) strategies for selecting the schools and/or districts to be monitored, (2) schedule for rotating the schools and/or districts to be monitored and (3) the training and roles of test administration monitors, as well as other key staff. If this responsibility is delegated to a regional center or district, include any memos or guidance issued by the state. Include examples of representative district(s) monitoring activities and results.
D. The extent of monitoring is likely to be evaluated on the basis of whether targeted monitoring is performed on the basis of past anomalies and whether samples of random monitoring are sufficient to ensure that if any given type of anomaly occurs in a substantial number of locations it is likely to be detected in at least some locations. Evidence that targeted and random monitoring are known by school and district staff to have occurred may be an important part of the evidence submitted to demonstrate that monitoring serves as a deterrent to anomalous test administration.

Critical Element 2.5 Test security

Critical Element	Examples of Evidence	Bite-sized Statements

The State has implemented and documented an appropriate set of policies and procedures to prevent test irregularities and ensure the integrity of test results through:
• Prevention of any assessment irregularities, including maintaining the security of test materials, proper test preparation guidelines and administration procedures, incident-reporting procedures, consequences for confirmed violations of test security, and requirements for annual training at the district and school levels for all individuals involved in test administration;
• Detection of test irregularities;
• Remediation following any test security incidents involving any of the State's assessments;
• Investigation of alleged or factual test irregularities.
	Collectively[A], evidence to support this critical element for the State's assessment system must demonstrate that the State has implemented and documented an appropriate approach to test security.
	
	Evidence to support this critical element for the State's assessment system may include:
	• State Test Security Handbook;
	• Summary results or reports of internal or independent monitoring, audit, or evaluation of the State's test security policies, procedures and practices, if any.
	
	Evidence of procedures for prevention of test irregularities includes documents such as:
	• Key documents, such as test coordinator manuals or test administration manuals for district and school staff, that include detailed security procedures for before, during and after test administration;
	• Documented procedures for tracking the chain of custody of secure materials and for maintaining the security of test materials at all stages, including distribution, storage, administration, and transfer of data;
	• Documented procedures for mitigating the likelihood of unauthorized communication, assistance, or recording of test materials (e.g., via technology such as smart phones);
	• Specific test security instructions for accommodations providers (e.g., readers, sign language interpreters, special education teachers and support staff if the assessment is administered individually), as applicable;
	• Documentation of established consequences for confirmed violations of test security, such as State law, State regulations or State Board-approved policies;
	• Key documents such as policy memos, listserv messages, test coordinator manuals and test administration manuals that document that the State communicates its test security policies, including consequences for violation, to all individuals involved in test administration;
	• Newsletters, listserv messages, test coordinator manuals, test administrator manuals and/or other key documents from the State that clearly state that annual test security training is required at the district and school levels for all staff involved in test administration;
	• Evidence submitted under Critical Element 2.3 – Test Administration that shows:
	  ○ The State's test administration training covers the relevant aspects of the State's test security policies;
	  ○ Procedures for ensuring that all individuals involved in test administration receive annual test security training.
	
	  For the State's technology-based assessments, evidence of procedures for prevention of test irregularities includes:
	• Documented policies and procedures for districts and schools to address secure test administration challenges related to hardware, software, internet connectivity, and internet access.
	
	Evidence of procedures for detection of test irregularities includes documents such as:
	• Documented incident-reporting procedures, such as a template and instructions for reporting test administration irregularities and security incidents for district, school and other personnel involved in test administration;
	• Documentation of the information the State routinely collects and analyzes for test security purposes, such as description of post-administration data forensics analysis the State conducts (e.g., unusual score gains or losses, similarity analyses, erasure/answer change analyses, pattern analysis, person fit analyses, local outlier detection, unusual timing patterns);
	• Summary of test security incidents from most recent year of test administration (e.g., types of incidents and frequency) and examples of how they were addressed, or other documentation that demonstrates that the State identifies, tracks, and
		1. Prevention of Test Irregularities
		a. Security procedures are in place for all who come into contact with test materials.[B]
		b. Procedures are in place to ensure the security of assessment materials before, during and after administration.[C]
		2. Detection of Test Irregularities
		a. Procedures for self-reporting or whistle-blowing are documented.
		b. Forensic analyses are performed to detect possible testing irregularities.[D,E]
		c. Procedures were carried out and results documented.
		3. Remediation and Investigation
		a. Investigative procedures for following-up on identified irregularities are clearly documented.
		b. Consequences for confirmed violations of test security are clearly documented.
		c. Clearly documented procedures and specifications are provided for responding to breaches in test security.[F]

Critical Element 2.5 Test security

Critical Element	Examples of Evidence	Bite-sized Statements

	resolves test irregularities.
	
	Evidence of procedures for remediation of test irregularities includes documents such as:
	• Contingency plan that demonstrates that the State has a plan for how to respond to test security incidents and that addresses:
	  ○ Different types of possible test security incidents (e.g., human, physical, electronic, or internet-related), including those that require immediate action (e.g., items exposed on-line during the testing window);
	  ○ Policies and procedures the State would use to address different types of test security incidents (e.g., continue vs. stop testing, retesting, replacing existing forms or items, excluding items from scoring, invalidating results);
	  ○ Communication strategies for communicating with districts, schools and others, as appropriate, for addressing active events.
	
	Evidence of procedures for investigation of alleged or factual test irregularities includes documents such as:
	• State's policies and procedures for responding to and investigating, where appropriate, alleged or actual security lapses and test irregularities that:
	  ○ Include securing evidence in cases where an investigation may be pursued;
	  ○ Include the State's decision rules for investigating potential test irregularities;
	  ○ Provide standard procedures and strategies for conducting investigations, including guidelines to districts, if applicable;
	  ○ Include policies and procedures to protect the privacy and professional reputation of all parties involved in an investigation.
	
	Note: Evidence should be redacted to protect personally identifiable information, as appropriate.[G]	

Potential Sources of Evidence to Support Bite-sized Statements

1. Test security handbook, test administration manuals[H], directions to proctors, directions to test takers, audit trails or other documentation that procedures were implemented, CAT exposure controls (i.e., methods used to minimize overexposure of any given item), test monitoring procedures, summary reports of test monitoring.
2. Procedures for self-reporting, procedures for whistle-blowing, summaries of self-reports, summaries of whistle-blowing irregularity reports, forensic analysis procedures[I], reports of forensic analyses.
3. Investigation procedures, training materials for investigators, procedures and criteria for issuing findings from investigations, procedures for responding to findings of investigations (including application of sanctions when appropriate), historical investigation reports.

Notes

A. A synthesis or executive summary of the evidence for the major parts of this critical element (prevention, detection, remediation, and investigation) should be provided to highlight the state's actions to ensure the integrity of test results for both the general and AA-AAS.
B. 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 materials can include:
   - 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.
   - **Computer Based Assessments:** training includes details necessary to support secure, online administration such as disabling "screenshot" abilities and web browsing.
C. For example:

Critical Element 2.5 Test security

Notes
• a documented chain of custody throughout the item and test development, review, and publishing cycle. It could also be shown that procedures are consistent with ED's 2013 Testing Integrity Symposium's Best Practices and other documents such as the CCSSO TILSA *Test Security Handbook*; International Test Commission *Guidelines for the Security of Examinations, Tests, and Other Assessments*; ATP/CCSSO *Operational Best Practices for Statewide Large Scale Assessment Programs*; and NCME *Testing and Data Integrity in the Administration of Statewide Assessment Programs*. • delivery of paper forms to LEAs in a manner that helps to prevent review prior to administration (e.g., shrink wrap). D. This information could include rules for flagging suspect cases, associated rationales, descriptions of analyses, evidence triangulation, and actions that are taken based upon the body of evidence. E. **Computer Based Assessments:** While erasure analyses are common for paper-based forms, a variety of different detection methods have been developed for computer-based assessments such as investigating answer changing behavior, login times and length, and flags for unusual response patterns or gains (e.g. unusual changes for the same grade/ content area for a specified unit, such as a class, school, district; or for a cohort of students such as a class or grade cohort across years.) F. Documented response actions could include: • identifying the source of the breach, • intervening if one or more items have been posted on the web, • analyses that will indicate the impact of the breach on scoring, • remediation actions for exposed items, and/or • if full forms are compromised, breach forms using the same test blueprint and psychometric requirements are readily available G. For more guidance regarding personally identifiable information (PII) please refer to element 2.6. H. The administration manual documents test security requirements and describes procedures for reporting deviations from administration procedures (i.e., suspected foul play). I. Neither the guidance nor the examples of evidence indicate that forensic analyses should be conducted by a third party. However third-party analyses could potentially be helpful to strengthen a submission in two ways: (1) to supplement expertise, and (2) to demonstrate independence.

Critical Element 2.6 – Systems for protecting data integrity & privacy

Critical Element	Examples of Evidence	Bite-sized Statements
The State has policies and procedures in place to protect the integrity and confidentiality of its test materials, test-related data, and personally identifiable information, specifically: • To protect the integrity of its test materials and related data in test development, administration, and storage and use of results; • To secure student-level assessment data and protect student privacy and confidentiality, including guidelines for districts and schools; • To protect personally identifiable information about any individual student in reporting, including defining the minimum number of students necessary to allow reporting of scores for all students and student groups.	Evidence to support this critical element for the State's general assessments and AA-AAS includes documents such as: • Evidence of policies and procedures to protect the integrity and confidentiality of test materials and test-related data, such as: ○ State security plan, or excerpts from the State's assessment contracts or other materials that show expectations, rules and procedures for reducing security threats and risks and protecting test materials and related data during item development, test construction, materials production, distribution, test administration, and scoring; ○ Description of security features for storage of test materials and related data (i.e., items, tests, student responses, and results); ○ Rules and procedures for secure transfer of student-level assessment data in and out of the State's data management and reporting systems; between authorized users (e.g., State, district and school personnel, and vendors); and at the local level (e.g., requirements for use of secure sites for accessing data, directions regarding the transfer of student data); ○ Policies and procedures for allowing only secure, authorized access to the State's student-level data files for the State, districts, schools, and others, as applicable (e.g., assessment consortia, vendors);[A] ○ Training requirements and materials for State staff, contractors and vendors, and others related to data integrity and appropriate handling of personally identifiable information;[B] ○ Policies and procedures to ensure that aggregate or de-identified data intended for public release do not inadvertently disclose any personally identifiable information; ○ Documentation that the above policies and procedures, as applicable, are clearly communicated to all relevant personnel (e.g., State staff, assessment, districts, and schools, and others, as applicable (e.g., assessment consortia, vendors)); ○ Rules and procedures for ensuring that data released by third parties (e.g., agency partners, vendors, external researchers) are reviewed for adherence to State Statistical Disclosure Limitation[C] (SDL) standards and do not reveal personally identifiable information. • Evidence of policies and procedures to protect personally identifiable information about any individual student in reporting, such as: ○ State operations manual or other documentation that clearly states the State's SDL rules for determining whether data are reported for a group of students or a student group, including: ▪ Defining the minimum number of students necessary to allow reporting of scores for a student group; ▪ Rules for applying complementary suppression (or other SDL methods) when one or more student groups are not reported because they fall below the minimum reporting size; ▪ Rules for not reporting results, regardless of the size of the student group, when reporting would reveal personally identifiable information (e.g., procedures for reporting "<10%" for proficient and above when no student scored at those levels);[D] ▪ Other rules to ensure that aggregate or de-identified data do not inadvertently disclose any personally identifiable information; ▪ State operations manual or other document that describes how the State's rules for protecting personally identifiable information are implemented.	**1. Security of Test materials** a. Procedures for maintaining and documenting item and test security are described for all stages of development, administration, scoring, and reporting.[E] b. These procedures are clearly communicated. c. Procedures were followed so that the security of test materials is maintained.[F] **2. Data Security** a. Procedures are in place to ensure all data are managed securely and access is strictly controlled. b. Procedures are defined for how data will be securely transferred[G] including within the State (e.g., during the administration and scoring process) or between the State and third parties (e.g., vendors). c. Procedures are defined for secure storage or destruction of data. d. These procedures are clearly communicated. **3. Student Privacy** a. Procedures are in place to ensure state compliance with Federal Educational Rights and Privacy Act (FERPA) and any additional state regulations related to maintaining student privacy. b. Procedures are in place to protect personally identifiable information from unauthorized access or use.[A] c. Procedures are in place to assure suppression of data with groups below a minimum size.[H] d. Guidelines are in place for LEAs to protect PII and are clearly communicated.

Critical Element 2.6 – Systems for protecting data integrity & privacy

Potential Sources of Evidence to Support Bite-sized Statements	
All bite-sized statements:	Data management and security procedures; procedures for secure item, test, and data storage and transfer; contract language regarding item, test, and data security and confidentiality; confidentiality and security agreements; state and contractor operations manuals; test security handbook; official communications with schools and districts; summaries of irregularity records.
3:	Cell suppression rules, state law regarding data privacy.

Notes

A. Access to PII is given on a "need-to-know" basis. It could also be shown that procedures are consistent with NCES' 2010 Data Stewardship Guidelines. Additional resources dealing with privacy can be found at NCES' Privacy Technical Assistance Center.
B. In addition to the training materials and documentation of the training requirements, evidence could include lists of staff who participated in the trainings and documentation of the effectiveness of training (e.g., evaluation forms).
C. This does not appear to be specifically defined. It appears to refer to the minimum cell size for which results may be publicly reported to avoid deductive disclosure of individually identifiable student data.
D. ED emphasizes that this an illustrative example of the type of evidence that can be supplied, but is not a requirement.
E. The evidence called for here includes but goes beyond that in element 2.5. Documentation should include control over item exposure beginning with item writing through administration, scoring, and item banking.
F. Potential evidence may include the following: audits, trainings, and/or confidentiality agreements
G. For digital transfers, approaches include the use of a secure file transfer protocol site or data encryption.
H. For example, subgroup reporting decisions could be shown to be consistent with the APA/AERA/NCME *standards* (see Standards 3.15 to 3.20, as well as Standard 13.5, on p. 70–72 and 211, respectively)
I. See evidence and notes related to element 2.4 for more detail regarding responses to deviation from test security procedures as well as responses to other breaches of test security.

Critical Element 3.1 – Overall validity, including validity based on content[A]

Critical Element	Examples of Evidence	Bite-sized Statements
The State has documented adequate overall validity evidence[B] for its assessments, and the State's validity evidence includes evidence that the State's assessments measure the knowledge and skills specified in the State's academic content standards, including: • Documentation of adequate alignment between the State's assessments and the academic content standards the assessments are designed to measure in terms of content (i.e., knowledge and process), the full range of the State's academic content standards, balance of content, and cognitive complexity; • If the State administers alternate assessments based on alternate academic achievement standards, the assessments show adequate linkage to the State's academic content standards in terms of content match (i.e., no unrelated content) and the breadth of content and cognitive complexity determined in test design to be appropriate for students with the most significant cognitive disabilities.	**Collectively**[B], across the State's assessments, evidence to support critical elements 3.1 through 3.4 for the State's general assessments and AA-AAS must document overall validity evidence generally consistent with expectations of current professional standards. Evidence to document adequate overall validity evidence for the State's general assessments and AA-AAS includes documents such as: • A chapter on validity in the technical report for the State's assessments that states the purposes of the assessments and intended interpretations and uses of results and shows validity evidence for the assessments that is generally consistent with expectations of current professional standards[C]; • Other validity evidence, in addition to that outlined in critical elements 3.1 through 3.4, that is necessary to document adequate validity evidence for the assessments. Evidence to document adequate validity evidence based on content for the State's general assessments includes: • Validity evidence based on the assessment content that shows levels of validity generally consistent with expectations of current professional standards, such as: ○ Test blueprints, as submitted under Critical Element 2.1—Test Design and Development; ○ A full form of the assessment in one grade for the general assessment in reading/language arts and mathematics (e.g., one form of the grade 5 mathematics assessment and one form of the grade 8 reading/language arts assessment);[1] ○ Logical or empirical analyses that show that the test content adequately represents the full range of the State's academic content standards [F]; ○ Report of expert judgment of the relationship between components of the assessment and the State's academic content standards; ○ Reports of analyses to demonstrate that the State's assessment content is appropriately related to the specific inferences made from test scores about student proficiency in the State's academic content standards for all student groups; • Evidence of alignment, including: ○ Report of results of an independent alignment study that is technically sound (i.e., method and process, appropriate units of analysis, clear criteria) and documents adequate alignment, specifically that: ▪ Each assessment is aligned to its test blueprint, and each blueprint is aligned to the full range of State's academic content standards; or ▪ Each assessment is aligned to the full range of the State's academic content standards, and the procedures the State follows to ensure such alignment during test development; ○ Description of a systematic process and timeline the State will implement to address any gaps or weaknesses identified in the alignment studies.	**1. Overall Validity** a. Validity evidence from various sources and of various types has been coherently synthesized to adequately support overall validity.[B] b. Each intended test score interpretation or use is clearly defined, and is supported by the synthesis of validity evidence.[D] **2. Alignment** a. Test blueprints are aligned to the state content standards[E, F] in terms of both the knowledge[G] and cognitive complexity[H], with respect to both overall coverage[I] and balance of coverage. [J] b. All test forms are reflective of the test blueprint, and are thus aligned to the state content standards.[K] c. The distribution of cognitive complexity for each test form and content area is sufficient to assess the depth and complexity of the state's standards (i.e., the assessment measures higher-order cognitive processes).[L] d. The distribution of item types for each test form and content area is sufficient to assess the depth and complexity of the standards being addressed.[M] e. **Computer Adaptive Assessments:** The combination of the item pool and item selection procedures assures that each student receives a test aligned to the blueprint and that

[1] The Department recognizes the need for a State to maintain the security of its test forms; a State that elects to submit a test form(s) as part of its assessment peer review submission should contact the Department so that arrangements can be made to ensure that the security of the materials is maintained. Such materials will be reviewed by the assessment peer reviewers in accordance with the State's test security requirements and agreements.

Critical Element 3.1 – Overall validity, including validity based on content[A]

Critical Element	Examples of Evidence	Bite-sized Statements
	 For the State's computer-adaptive general assessments: • Empirical evidence that the size of the item pool and the characteristics (non-statistical (e.g., content) and statistical) of items it contains are appropriate for the test design and adequately reflect the blueprint in terms of: ○ Full range of the State's grade-level academic content standards; ○ Balance of content; ○ Cognitive complexity for each standard tested; ○ Range of item difficulty levels for each standard tested; ○ Structure of the assessment (e.g., number of items and proportion of item and response types specified by the blueprints); ○ Item pool size and composition sufficient to avoid over-exposure of items; • Results of an alignment study confirming that the test forms generated for individual students are aligned to the State's academic content standards in terms of: ○ Full range of the State's grade-level academic content standards; ○ Balance of content; ○ Cognitive complexity for each standard tested; ○ Range of item difficulty levels for each standard tested; ○ Structure of the assessment (i.e., features specified in Critical Element 2.1 – Test Design and Development, such as number of items and proportion of item and response types specified by the blueprints); • Empirical analyses that show: ○ The actual test forms produce an adequately precise estimate of student achievement; ○ Students are appropriately routed to the next item or stage based on their responses to the previous item or stage; ○ Response data adequately fit the psychometric model selected by the State. **AA-AAS.** For the State's AA-AAS, evidence to document adequate validity evidence based on content includes: • Validity evidence that shows levels of validity generally considered adequate by professional judgment regarding such assessments, such as: ○ Test blueprints and other evidence submitted under Critical Element 2.1 – Test Design and Development; ○ Evidence documenting adequate linkage between the assessments and the academic content they are intended to measure; ○ Other documentation that shows the State's assessments measure only the knowledge and skills specified in the State's academic content standards (or extended academic content standards, as applicable) for the tested grade (i.e., not unrelated content); • Evidence of alignment, such as: ○ Report of results of an independent alignment study that is technically sound and document adequate linkage between each of the State's assessments and the academic content the	assesses the depth and breadth of the state content standards.[N] f. **Computer Adaptive Assessments:** The combination of the item pool and item selection procedures assures that each student receives a test with adequate precision. g. **AA-AAS:** There is a clear and documented link between the general standards and the alternate standards, with a sound rationale for reduction in depth, breadth, and/or complexity; and that the alternate content standards retain the maximum complexity appropriate for the AA-AAS population.[O] h. **AA-AAS:** The usability and appropriateness of the test mode and test items (including accessibility features) are adequate.[P] i. **AA-AAS:** There is a clearly defined, adequate link between the assessments and the academic content standards or extended academic content standards. If the State has created extended academic content standards, they are appropriately linked to the academic content standards. j. A process and timeline for addressing issues identified with alignment is in place.

Critical Element 3.1 – Overall validity, including validity based on content[A]

Critical Element	Examples of Evidence	Bite-sized Statements
	assessments are designed to measure; ○ If the State developed extended academic content standards for students with the most significant cognitive disabilities and used these to develop its AA-AAS, the alignment study should document the linkage between the State's academic content standards and extended academic content standards as well as adequate linkage between the extended academic content standards and the assessments; • For an adaptive AA-AAS: ○ Summary of an analysis to confirm that the item pool adequately represents the test blueprints, such as a crosswalk of the item pool and the test blueprints; ○ Results of an alignment study that confirm that the test design, as implemented, produces assessments with adequate linkage to the academic content standards the assessments are designed to measure.	

Potential Sources of Evidence to Support Bite-sized Statements

1. All critical elements related to test validity, other documents related to other aspects of validity.
2. Legislation, regulation, and rule; test development procedures and specifications/ blueprints; technical manuals; procedure documents; item specifications; test specifications; test blueprints; alignment studies; official statements of intended assessment purposes and uses; CAT administration rules, item selection procedures, simulation results, and item pool sufficiency studies; alternate content standards development process.

Notes

A. Overlaps strongly with critical element 2.1.
B. Evidence gathered for the specific parts of element 3.1 and for elements 3.2–3.4 alone is insufficient. All evidence must be synthesized, and may be supplemented as needed: Evidence should be integrated into a "coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses" (APA/AERA/NCME, 2014, p. 21). This account is often presented as a validity argument. ED does not require consequential validity evidence in the overall validity argument. However, if available, such evidence can strengthen an overall validity argument. Older assessment programs will be in a better position to have such evidence.
C. APA/AERA/NCME *Standards* details five types of validity evidence (p. 13–19), the first four of which are also the types of evidence elicited under elements 3.1 to 3.4 (evidence based on test content, response processes, internal structure, and relations to other variables). As noted above, the S*tandards* also explain that validity evidence is meant to be integrated into an argument for the validity of the intended test score interpretations (p. 21–22).
D. The goal is to argue that intended interpretations and uses are well supported and defensible (APA/AERA/NCME). All intended interpretations and uses should be included, which include specific uses/interpretations in at least the following categories: student promotion and graduation decisions, teacher and school evaluation (including federally and/or state mandated school accountability decisions), and evaluation of instructional programs or materials. One way to capture the logic underlying the use of assessment results is through in a theory of action or interpretive argument. This documentation generally outlines the range of assumptions that must hold in order for the results to be interpreted in the intended manner and for the assessment to have the desired impact on systems and stakeholders. This documentation may take a variety of different formats and may 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.
E. If speaking and/or listening are included in state content standards but not in state assessment, a waiver will need to submitted to ED. Details of the waiver have not been released, but it is known that the waiver will require a plan for speaking and listening to be incorporated in state assessment in the future.
F. For states administering common assessments (e.g., consortia tests, ACT), if the assessments were developed to measure a specific set of content standards and the state has supplemented that with state-specific standards, the state will need to provide evidence to show how they have attended to that difference in state assessment.
G. Categories of content knowledge in the content standards.
H. Levels of cognitive complexity in the standards (typically using discipline-specific Depth of Knowledge or another taxonomy of levels of complexity in cognitive processes).

Critical Element 3.1 – Overall validity, including validity based on content[A]

Notes
I. Test blueprints reflect coverage of the full range of the content standards (not necessarily that every content standard is represented by at least one item, but that the blueprint reflects the full range of the standards). J. Test blueprint is balanced in representing different parts of the content standards (e.g., no substantial over- or under-representation of subsets of content standards). K. Could be based on studies of the alignment of individual items on various forms to the intended blueprint or to the content standards themselves. It may not be necessary to study alignment of every form in every grade and subject, but to demonstrate that the processes for item and test form development leads to the administration of aligned tests. L. Although this is addressed implicitly as part of previous *Bite-sized Statements*, the wording of the *Guidance* suggests that this specific aspect should be explicitly and separately addressed. However, this statement does not imply that performance tasks are required. ED has indicated that the important evaluation is that the assessment measures the full range (depth & breadth) of the state's adopted standards. M. Although appropriate use of item types to achieve alignment is addressed implicitly in previous *Bite-sized Statements*, the wording of the *Guidance* implies that this should be explicitly and separately addressed. N. Requires that all students experience a test consistent with the same blueprint, even though no student may receive the same test as any other. It is important to demonstrate that the item selection procedures prioritize meeting the blueprint and achieving alignment. O. Much of the evidence for this *Bite-sized Statement* can be drawn from evidence developed for critical elements 1.1 and 1.2. P. Accessibility considerations should be woven throughout the item and test development process so that the tests are accessible to the highest number of students. Issues of accessibility should be discussed from test design through test administration.

Critical Element 3.2 – Validity based on cognitive processes

Critical Element	Examples of Evidence	Bite-sized Statements
The State has documented adequate validity evidence that its assessments tap the intended cognitive processes appropriate for each grade level as represented in the State's academic content standards[A].	Evidence to support this critical element for the State's general assessments includes: • Validity evidence based on cognitive processes that shows levels of validity generally consistent with expectations of current professional standards, such as: ○ Results of cognitive labs[B] exploring student performance on items that show the items require complex demonstrations or applications of knowledge and skills; ○ Reports of expert judgment of items that show the items require complex demonstrations or applications of knowledge and skills; ○ Empirical evidence that shows the relationships of items intended to require complex demonstrations or applications of knowledge and skills to other measures that require similar levels of cognitive complexity in the content area (e.g., teacher ratings of student performance, student performance on performance tasks or external assessments of the same knowledge and skills). **AA-AAS.** For the State's AA-AAS, evidence to support this critical element includes[C]: • Validity evidence that shows levels of validity generally considered adequate by professional judgment regarding such assessments, such as: ○ Results of cognitive labs exploring student performance on items that show the items require demonstrations or applications of knowledge and skills; ○ Reports of expert judgment of items that show the items require demonstrations or applications of knowledge and skills; ○ Empirical evidence that shows the relationships of items intended to require demonstrations or applications of knowledge and skills to other measures that require similar levels of cognitive complexity in the content area (e.g., teacher ratings of student performance, student performance on performance tasks or external assessments of the same knowledge and skills).	Not needed.

Potential Sources of Evidence to Support Bite-sized Statements
Cognitive laboratory procedures, cognitive laboratory reports, expert judgment reports such as educator judgments of items' depth of knowledge level, test design and development processes, reports of external relationships of items to similar judgments of cognitive complexity, item and task specifications or templates, learning progressions or learning maps, item review training, item review reports, and surveys regarding the students' experience taking the assessment.

Notes

A. Intended cognitive processes are often defined through Webb's Depth of Knowledge levels in conjunction with the academic content standards (and captured in each assessment's blueprint). Other approaches to defining intended cognitive processes include learning progressions, learning maps or design patterns. Items leased or purchased from vendors and/or existing item banks and aligned to new content standards will likely require vendors or states to demonstrate of alignment to the intended cognitive processes of the new standards.

B. Cognitive labs often take the form of "Think Alouds," in which students are asked to explain their thinking. This helps identify the strategies students use to solve the problems so the alignment of the cognitive processes used with the intended content standards can be evaluated. Examples of cognitive labs within the context of education assessment include work on NAEP's background indicators and PISA's first phase of item development. Given the cost of cognitive labs, it will not be possible to explore every item – thus cognitive labs will need to target to a sample of items. These samples could be selected to be representative of the total item pool or strategically focused (e.g., items thought to elicit higher order thinking or with novel formats).

C. Does not appear to require anything above that required for general assessments.

Critical Element 3.3 – Validity based on internal structure

Critical Element	Examples of Evidence	Bite-sized Statements

The State has documented adequate validity evidence that the scoring and reporting structures of its assessments are consistent with the sub-domain structures of the State's academic content standards on which the intended interpretations and uses of results are based[A].	Evidence to support this critical element for the State's general assessments includes:
	• Validity evidence based on the internal structure[B] of the assessments that shows levels of validity generally consistent with expectations of current professional standards[C], such as:
	  ○ Reports of analyses of the internal structure of the assessments (e.g., tables of item correlations) that show the extent to which the interrelationships among subscores are consistent with the State's academic content standards for relevant student groups;
	  ○ Reports of analyses that show the dimensionality of the assessment is consistent with the structure of the State's academic content standards and the intended interpretations of results; [D]
	  ○ Evidence that ancillary constructs needed for success on the assessments do not provide inappropriate barriers for measuring the achievement of all students, such as evidence from cognitive labs or documentation of item development procedures; [E]
	  ○ Reports of differential item functioning (DIF) analyses that show whether particular items (e.g., essays, performance tasks, or items requiring specific knowledge or skills) function differently for relevant student groups.
	
	AA-AAS. For the State's AA-AAS, evidence to support this critical element includes:
	• Validity evidence that shows levels of validity generally considered adequate by professional judgment regarding such assessments, such as:
	  ○ Validity evidence based on the internal structure of the assessments, such as analysis of response patterns for administered items (e.g., student responses indicating no attempts at answering questions or suggesting guessing).
		1. Internal structure
		a. If subscores are reported, their structure is consistent with the structure of the content standards[D].
		b. The internal structure of assessments is consistent with the structure of the content standards.[D,F]
		c. The internal structure of each assessment is linked to the intended test score interpretation(s).[G]
		2. Construct irrelevant variance
		a. Internal structure is similar across subgroups
		b. Construct irrelevant variance is minimized.[E]
		c. **AA-AAS**: There is minimal evidence of inattentiveness and guessing.[H]

Potential Sources of Evidence to Support Bite-sized Statements

1. Content standards, content specifications, test blueprints, sample reports (with PII redacted), intended uses and interpretations, item correlations, dimensionality analyses, item analyses.
2. Measurement invariance analyses, item analyses, DIF analyses, item and person fit analyses.

Notes

A. It is unclear whether this calls for subscore reporting. However, if subscores are reported, the structure of the standards and subscore reporting should be consistent.
B. The internal structure of an assessment can be examined in a variety of ways, commonly including simple bivariate correlations, exploratory factor analysis and related data reduction techniques such as principle components analysis, confirmatory factor analysis, or structural equation modeling.
C. Standards 1.13 to 1.15 of the APA/AERA/NCME *Standards* (p. 26–27, see also p. 16) address internal structure.
D. For example, a single score might be reported. Analysis showing essential unidimensionality or a rationale for the creation of a composite score would support this type of reporting. Alternatively, multiple scores (including subscores) could be reported, supported by analyses showing that dimensionality aligns reasonably to these multiple reported scores (e.g., model-data fit to a multidimensional scoring model, scores are sufficiently distinct from one another to support separate inferences).
E. Possibly shown through cognitive labs, reliability analyses, item development training, item submission certifications, bias and sensitivity review procedures, and/or DIF analyses.
F. For example, finding that item correlations within each content standard domain are stronger than across domains would provide evidence of consistency between the assessment's internal structure and the standards.
G. For example, construct-irrelevant language complexity does not interfere with performance on mathematics items, the specificity of scores (e.g., overall vs. overall and subscores) is consistent with intended uses.
H. Typically established through item/person fit analyses.

Critical Element 3.4 – Validity based on relations to other variables

Critical Element	Examples of Evidence	Bite-sized Statements
The State has documented adequate validity evidence that the State's assessment scores[A] are related as expected[B] with other variables.	Evidence to support this critical element for the State's general assessments includes: • Validity evidence that shows the State's assessment scores are related as expected with criterion and other variables for all student groups, such as: ○ Reports of analyses that demonstrate positive correlations between State assessment results and external measures that assess similar constructs, such as NAEP, TIMSS, assessments of the same content area administered by some or all districts in the State, and college-readiness assessments; ○ Reports of analyses that demonstrate convergent relationships between State assessment results and measures other than test scores, such as performance criteria, including college- and career-readiness (e.g., college-enrollment rates; success in related entry-level, college credit-bearing courses; post-secondary employment in jobs that pay living wages)[C]; ○ Reports of analyses that demonstrate positive correlations between State assessment results and other variables, such as academic characteristic of test takers (e.g., average weekly hours spent on homework, number of advanced courses taken)[D]; ○ Reports of analyses that show stronger positive relationships with measures of the same construct than with measures of different constructs; ○ Reports of analyses that show assessment scores at tested grades are positively correlated with teacher judgments of student readiness at entry in the next grade level. **AA-AAS.** For the State's AA-AAS, evidence to support this critical element includes: • Validity evidence that shows levels of validity generally considered adequate by professional judgment regarding such assessments, such as: ○ Validity evidence based on relationships with other variables, such as analyses that demonstrate positive correlations between assessment results and other variables, for example: ▪ Correlations between assessment results and variables related to test takers (e.g., instructional time on content based on grade-level content standards); ▪ Correlations between proficiency on the high-school assessments and performance in post-secondary education, vocational training or employment.[E]	**1. Relationships to Other Variables** a. Test scores[A] are more[F] strongly related to scores on other assessments of similar constructs. b. Test scores are more[G] weakly related to scores on assessments of dissimilar constructs[H]. c. Test scores are positively related to other relevant criteria not based on standardized assessment, including (1) empirical indicators of career- and college-readiness/success and (2) indicators of academic performance in school.[I]

Potential Sources of Evidence to Support Bite-sized Statements[J]

Plans for and/or results of correlational studies of test scores and other test scores, plans for and/or results of correlational studies of test scores and non-test-based indicators, plans for and/or results of studies of the relationships between test scores and related later outcomes, plans for and/or results of distributional comparisons (e.g., of state test scores/performance levels to NAEP, AP, ACT, SAT, college enrollment, college remediation, and college grade outcomes)[K], and analyses of test scores and achievement gaps by demographics and prior achievement[L].

Notes

A. For some criteria, it is the performance level classifications that will need to be investigated rather than the scale scores themselves. Thus, for this element, "assessment scores" may mean the scale scores or the performance level classifications created based on the scale scores.
B. It is likely that stating the direction of the expected relationship will be sufficient for most variables. However, when historical data are available the expectations can be made more specific. Consider the following hypothetical example – "given that scores on prior high school mathematics assessments correlated with high school grade point averages at 0.5, we expect to find a positive correlation of similar magnitude."
C. Given the recent emphasis on career and college readiness, the relationships between test scores and empirical (not predictive) indicators of college- and career-readiness are likely to be particularly important for assessments in the upper grades. However, for recently administered assessments, data on readiness will not yet be available. Some

Critical Element 3.4 – Validity based on relations to other variables

Notes
comparisons can still be made. For example, the percentage of students classified as college- and career-ready by the assessments and the historical percentages of students enrolling in college could be compared to provide a rough indication of the relationship between the test score and academic career and college readiness. Additionally, teacher judgements of college and career readiness could be used. However, plans for gathering indicators from post-secondary education should be included. ***For newer assessment programs, this evidence will not likely exist***. A rationale for predicted post-secondary education, vocational training or employment should be included, as should plans for collecting longitudinal data to support analyses of these relationships should be detailed. To create an alternate set of data in advance of that timeline, the high-school test could be administered to college freshman.
D. In many states, such additional data (e.g., course-taking behavior, grades, time spent on homework) may not be available to the state (i.e., the data may be owned by districts). This may require agreements between districts and the state (if possible), surveys of a sample of students (if possible), and/or a description of why obtaining such data is not feasible.
E. For newer assessment programs, this evidence will not likely exist. A rationale and definition should be given for what success in post-secondary education, vocational training, and/or employment means for the AA-AAS population should be provided. If evidence has not already been gathered, plans for collecting longitudinal data to analyze the relationship between high-school AA-AAS scores and post-secondary outcomes should be detailed.
F. Compared to assessments of dissimilar constructs.
G. Compared to assessments of similar constructs.
H. These relationships are often called sources of "discriminant evidence" (see AERA/APA/NCME *Standards* p. 16–17 and 27–30). Consider a hypothetical example of a State's 4th grade mathematics assessment. The expectation is that scores on that test would be more strongly (e.g., positively correlated) with other measures dealing with mathematical achievement content taught near the 4th grade (e.g., 3rd or 5th grade state test mathematics scores) than measures dealing with, say, English language arts achievement like that State's 4th grade ELA test.
I. These indicators could include student's school grade point average or subject specific grades, remediation or retention decisions (e.g., teacher judgments of student readiness for the next grade), average weekly hours spent on homework, or number of advanced courses taken.
J. The collective evidence should be synthesized to demonstrate that test scores are more strongly related with similar external criteria (whether they are test-based or not) and less strongly related with dissimilar external criteria (whether they are test-based or not).
K. For example, to demonstrate (as described in Note C) that proficiency rates are consistent with rates of related outcomes, even if they are based on different samples.
L. Analyses that show that the relationship between demographics and test scores is reduced when controlling for prior achievement (however measured) can be helpful in identifying that test scores are more highly correlated with school-based variables (prior achievement) than with non-school-based variables. Similar analyses showing that achievement gaps (between a focal and comparison group) are smaller for students with similar prior achievement than for all students in both groups can also be helpful in identifying that gaps in achievement are reduced when controlling for school-based measures.

Critical Element 4.1 - Reliability

Critical Element	Examples of Evidence	Bite-sized Statements

The State has documented adequate reliability evidence[A] for its assessments for the following measures of reliability for the State's student population overall and each student group and, if the State's assessments are implemented in multiple States, for the assessment overall and each student group, including:
• Test reliability of the State's assessments estimated for its student population;
• Overall and conditional standard error of measurement of the State's assessments;
• Consistency and accuracy of estimates in categorical classification decisions for the cut scores and achievement levels based on the assessment results;
 	
For computer-adaptive tests, evidence that the assessments produce test forms with adequately precise estimates of a student's achievement.
	Collectively[B], evidence for the State's general assessments and AA-AAS must document adequate reliability[C] evidence generally consistent with expectations of current professional standards.
	
	Evidence to support this critical element for the State's general assessments includes documentation such as:
	• A chapter on reliability in the technical report for the State's assessments that shows reliability evidence;
	• For the State's general assessments, documentation of reliability evidence generally consistent with expectations of current professional standards, including:
	  ○ Results of analyses for alternate-form or, test-retest internal consistency reliability statistics, as appropriate, for each assessment;
	  ○ Report of standard errors of measurement and conditional standard errors of measurement, for example, in terms of one or more coefficients or IRT-based test information functions at each cut score specified in the State's academic achievement standards;
	  ○ Results of estimates of decision consistency and accuracy for the categorical decisions (e.g., classification of proficiency levels) based on the results of the assessments.
	
	For the State's computer-adaptive assessments, evidence that estimates of student achievement are adequately precise includes documentation such as:
	• Summary of empirical analyses showing that the estimates of student achievement are adequately precise for the intended interpretations and uses of the student's assessment score;
	• Summary of analyses that demonstrates that the test forms are adequately precise across all levels of ability in the student population overall and for each student group (e.g., analyses of the test information functions and conditional standard errors of measurement).
	
	AA-AAS. For the State's AA-AAS, evidence to support this critical element includes:
	• Reliability evidence that shows levels of reliability generally considered adequate by professional judgment regarding such assessments includes documentation such as:
	  ○ Internal consistency coefficients that show that item scores are related to a student's overall score;
	  ○ Correlations of item responses to student proficiency level classifications;
	  ○ Generalizability evidence such as evidence of fidelity of administration;
	  ○ As appropriate and feasible given the size of the tested population, other reliability evidence as outlined above.
		1. Score Reliability/Precision
		a. Each assessment's scores have an adequate level of reliability[D] for the State's population, both overall and for accountability subgroups.
		b. Each assessment's scores have sufficiently small conditional standard errors of measurement[E], both overall and for accountability subgroups.
		c. Each assessment's proficiency level classifications are adequately consistent and accurate[F], both overall and for accountability subgroups.
		d. **Computer-Adaptive Assessment**: Estimates of student achievement produced by the adaptive forms have sufficiently small standard errors of measurement (i.e., conditional standard errors), both overall and for accountability subgroups.
		e. Evaluation of the adequacy of reliability/precision is keyed to the intended uses and purposes of assessment scores.
		f. **AA-AAS**: Item responses have strong positive associations[G] with student overall scale scores and proficiency level classifications.
		g. **AA-AAS:** Assessments were administered with fidelity to administration instructions.

Potential Sources of Evidence to Support Bite-sized Statements
• Technical reports, supplemental analyses of conditional standard errors and/or classification accuracy or consistency, item-total correlations, proficiency /achievement level reports, tables of scale scores and conditional standard errors of measurement, generalizability studies, evaluations of fidelity of administration to instructions.

Critical Element 4.1 - Reliability

Notes
A. Evidence of reliability will need to be presented for all reported scores (e.g., scale scores, subscores, growth scores). Standard 2.3 (APA/AERA/NCME, 2014, p. 43) notes that "For each total score, subscores, or combination of individual scores that is to be interpreted, estimates of relevant indices of reliability/precisions should be reported). Note that in many cases, reliability (increased by maximizing similarity of items) and validity (maximized by including a variety of item formats measuring achievement differently) may be in conflict. Chasing high reliability at the expense of validity should be avoided, but reliability should be reasonably high.
B. The evidence of reliability must be summarized and synthesized. The evidence alone is insufficient. The case must be made that the assessment scores and related classifications are reliable or precise enough support the intended test score interpretations detailed under critical element 3.1.
C. Reliability here refers to "the more general notion of consistency of the scores across instances of the testing procedure" (APA/ AERA/NCME, 2014, p. 33) instead of just a reliability coefficient from classical test theory.
D. Standard 2.5 (APA/AERA/NCME, 2014, p. 46) notes that "when there is credible evidence for expecting that conditional standard errors of measurement or test information functions will differ substantially for various subgroups, investigation of the extent and impact of such differences should be undertaken and reported as soon as is feasible."
E. Reliability/Precision of scale scores could be demonstrated through Classical Test Theory statistics, like the internal consistency statistic Cronbach's alpha, or Item Response Theory, statistics like marginal reliability or average conditional standard error.
F. Potentially quantified by indices of classification consistency and accuracy (see, for example, p. 99–101 of the 4th Edition of *Educational Measurement*) or examinations of the amount of information (shown through an IRT information function) around each cut point.
G. Potentially shown using an item-total correlation or average item score by proficiency classification.

Critical Element 4.2 – Fairness & accessibility

Critical Element	Examples of Evidence	Bite-sized Statements
The State has taken reasonable and appropriate steps to ensure that its assessments are accessible to all students and fair across student groups in the design, development and analysis of its assessments.	Evidence to support this critical element for the State's general assessments and AA-AAS includes: For the State's general assessments: • Documentation of steps the State has taken in the design and development of its assessments, such as: ○ Documentation describing approaches used in the design and development of the State's assessments (e.g., principles of universal design, language simplification, accessibility tools and features embedded in test items or available as an accompaniment to the items); ○ Documentation of the approaches used for developing items; ○ Documentation of procedures used for maximizing accessibility of items during the development process, such as guidelines for accessibility and accessibility tools and features included in item specifications; ○ Description or examples of instructions provided to item writers and reviewers that address writing accessible items, available accessibility tools and features, and reviewing items for accessibility; ○ Documentation of procedures for developing and reviewing items in alternative formats or substitute items and for ensuring these items conform with item specifications; ○ Documentation of routine bias and sensitivity training for item writers and reviewers; ○ Documentation that experts in the assessment of students with disabilities, English learners and individuals familiar with the needs of other student populations in the State were involved in item development and review; ○ Descriptions of the processes used to write, review, and evaluate items for bias and sensitivity; ○ Description of processes to evaluate items for bias during pilot and field testing; ○ Evidence submitted under Critical Elements 2.1 – Test Design and Development and Critical Element 2.2 – Item Development; • Documentation of steps the State has taken in the analysis of its assessments, such as results of empirical analyses (e.g., DIF and differential test functioning (DTF) analyses) that identify possible bias or inconsistent interpretations of results across student groups. **AA-AAS.** For the State's AA-AAS: • Documentation of steps the State has taken in the design and development of its assessments, as listed above; • Documentation of steps the State has taken in the analysis of its assessments, for example: ○ Results of bias reviews or, when feasible given the size of the tested student population, empirical analyses (e.g., DIF analyses and DTF analyses by disability category[C]); ○ Frequency distributions of the tested population by disability category[C]; ○ As appropriate, applicable and feasible given the size of the tested population, other evidence as outlined above. Note: This critical element is closely related to Critical Element 2.2 – Item Development.	**1. Item and Stimulus[A] Development** a. Item/stimulus writers received sound training on bias, accessibility, and fairness (including incorporation of accessibility tools). b. Procedures and criteria used to develop fair and accessible items were clearly documented and followed. c. Experts for various demographic populations were included. **2. Item and Stimulus Review** a. Sound procedures and criteria were used to review items for bias, fairness and accessibility were clearly documented and followed. b. Item reviewers received sound training on bias, inaccessibility, and unfairness. c. Incorporates expert judgment and quantitative analyses[B] of field test data. d. Experts for various demographic populations were included. e. **AA-AAS**: Bias, fairness and accessibility were investigated and documented for students eligible to take the AA-AAS. **3. Accessibility Tools** a. Procedures and criteria used to develop accessibility tools are documented.[D]

Critical Element 4.2 – Fairness & accessibility

Potential Sources of Evidence to Support Bite-sized Statements
1 & 2: Item and stimulus writing training materials; procedures and criteria for submitting and accepting items and stimuli; audit trails for procedures and criteria; technical manuals; and procedures and results of quantitative analyses of potential bias, sensitivity, unfairness or inaccessibility. 3: Test specifications, accommodations manual.

Notes
A. Includes such content as passages, artwork, photographs, charts, and tables that students must refer to in order to respond to test items. B. Such as differential item function, differential test functioning, differential scoring engine functioning. C. This appears to be new as an interpretation of what the critical element means, and it will require considerable thought and effort to analyze. If a state has determined that specific disability is not a critical data element and does not require districts to submit such data, the rationale for that decision will likely be important to include. D. While the critical element and examples of evidence do not explicitly call for demonstration that accessibility tools were implemented and made available, documentation of implementation and availability is likely to strengthen a peer review submission. Documenting the frequency of use (if possible) will also likely add to the strength of a submission.

Critical Element 4.3 – Full performance continuum

Critical Element	Examples of Evidence	Bite-sized Statements
The State has ensured that each assessment provides an adequately precise estimate of student performance across the full performance continuum, including for high- and low-achieving students[A].	Evidence to support this critical element for the State's general assessments and AA-AAS includes documents such as: For the State's general assessments: • Description of the distribution of cognitive complexity and item difficulty indices that demonstrate the items included in each assessment adequately cover the full performance continuum; • Analysis of test information functions (TIF) and ability estimates for students at different performance levels across the full performance continuum or a pool information function across the full performance continuum; • Table of conditional standard errors of measurement at various points along the score range. **AA-AAS.** For the State's AA-AAS: • A cumulative frequency distribution or histogram of student scores for each grade and subject on the most recent administration of the State's assessment[B]; • For students at the lowest end of the performance continuum (e.g., pre-symbolic language users or students with no consistent communicative competencies), evidence that the assessments provide appropriate performance information (e.g., communicative competence); • As appropriate, applicable and feasible given the size of the tested population, other evidence as outlined above.	**1. Precision Across the Full Performance Continuum** a. Test scores are adequately precise across the entire performance continuum (including the extremes).[C] b. **CAT**: The range of item difficulties is not unreasonably restricted based on level of cognitive complexity.[D] c. **AA-AAS**: Assessments produce adequate information for student at the lowest levels of performance[B].

Potential Sources of Evidence to Support Bite-sized Statements

Tables or graphs of scale scores by conditional standard errors of measurement, tables or graphs of scale scores by test information, item and task specifications, test blueprints, histograms or frequency distributions of observed scale scores[B].

Notes

A. This critical element appears to be designed to demonstrate two concepts: (1) ceiling and floor effects have been avoided, and (2) conditional standard errors of measurement are not too large for the intended uses and purposes for students on the extremes of the performance continuum.
B. To demonstrate lack of ceiling and/or floor effects.
C. Could be shown directly from a comparison of item difficulties to student performance estimates (particularly in the case of the Rasch and related one parameter IRT models) or through functions of item difficulty, such as test information functions.
D. For CAT, there should be a range of item difficulties for all levels of cognitive complexity, although overlap between those ranges can be expected to be limited.

Critical Element 4.4 - Scoring

Critical Element	Examples of Evidence	Bite-sized Statements

The State has established and documented standardized scoring procedures and protocols for its assessments that are designed to produce reliable results, facilitate valid score interpretations, and report assessment results in terms of the State's academic achievement standards.	Evidence to support this critical element for the State's general assessments and AA-AAS includes:
	• A chapter on scoring in a technical report for the assessments or other documentation that describes scoring procedures, including:
	  ○ Procedures for constructing scales used for reporting scores and the rationale for these procedures;
	  ○ Scale, measurement error, and descriptions of test scores;
	• For scoring involving human judgment:
	  ○ Evidence that the scoring of constructed-response items and performance tasks includes adequate procedures and criteria for ensuring and documenting inter-rater reliability (e.g., clear scoring rubrics, adequate training for and qualifying of raters, evaluation of inter-rater reliability, and documentation of quality control procedures);
	  ○ Results of inter-rater reliability of scores on constructed-response items and performance tasks;
	• For machine scoring of constructed-response items:
	  ○ Evidence that the scoring algorithm and procedures are appropriate, such as descriptions of development and calibration, validation procedures, monitoring, and quality control procedures;
	  ○ Evidence that machine scoring produces scores are comparable to those produced by human scorers, such as rater agreement rates for human- and machine-scored samples of responses (e.g., by student characteristics such as varying achievement levels and student groups), systematic audits and rescores;
	• Documentation that the system produces student results in terms of the State's academic achievement standards;
	• Documentation that the State has rules for invalidating test results when necessary (e.g., non-attempt, cheating, unauthorized accommodation or modification) and appropriate procedures for implementing these rules (e.g., operations manual for the State's assessment and accountability systems, test coordinators manuals and test administrator manuals, or technical reports for the assessments)[A].
	
	AA-AAS. For the State's AA-AAS, in addition to the above:
	• If the assessments are portfolio assessments, evidence of procedures to ensure that only student work including content linked to the State's grade-level academic content standards is scored;
	• If the alternate assessments involve any scoring of performance tasks by test administrators (e.g., teachers):
	  ○ Evidence of adequate training for all test administrators (may include evidence submitted under Critical Element 2.3 – Test Administration);
	  ○ Procedures the State uses for each test administration to ensure the reliability of scoring;
	  ○ Documentation of the inter-rater reliability of scoring by test administrators.
		1. Human Scoring
		a. Scoring procedures and criteria are clearly documented and were followed.
		b. Procedures and criteria for evaluating inter-rater reliability are well documented and provide for reliable results.
		c. Validity evidence supports the intended use and interpretation of scores resulting from human judgments[B].
		d. **AA-AAS**: For tasks scored by test administrators, administrators had adequate training on appropriate procedures and scoring criteria.
		2. Machine Scoring
		a. Scoring procedures are clearly documented and were followed.
		b. Validity evidence supports the intended use and interpretation of machine scored results.[C]
		c. The reliability (consistency) of machine scoring is supported.[D]
		3. Item Scores, Scale Scores, and Proficiency Levels
		a. Item scoring rules were correctly applied.
		b. Collections of item responses have been correctly converted to scale scores.
		c. Scale scores have been correctly converted to proficiency levels.
		d. Proficiency levels are reported based on officially adopted cut scores[E]
		e. Invalidation procedures (e.g., for irregularities that render scores invalid) are clearly described, were clearly communicated to schools and districts, and were followed.

Critical Element 4.4 - Scoring

Potential Sources of Evidence to Support Bite-sized Statements
1. Technical reports; human scoring training materials, inter-rater reliability reports, validity reports (e.g., seeded validity papers); human rater qualification procedures, criteria, and reports; human rater monitoring, retraining, dismissal, and other quality control procedures, criteria, and reports; human scoring procedures (including resolution and rescoring as needed), audit trails for compliance with procedures and criteria. 2. Technical reports, machine scoring guidelines, technical specifications, and criteria; machine scoring reliability and validity reports; machine scoring quality control procedures; human read-behind procedures, criteria, and reports; aberrant response detection and routing procedures, criteria, and reports (i.e., to handle responses not well scored by machine); audit trails for compliance with procedures and criteria. 3. Technical reports, item scoring procedures, criteria, and reports[E], raw-score-to-scale score conversion tables, diagnostic output from scoring software, scale score to proficiency level tables, quality control procedures, audit trails and audit reports.

Notes
A. If states do not have sole authority to invalidate test scores, it is likely that documentation will be needed that the entity with responsibility (or entities with joint responsibility) for invalidating test scores meet this requirement. B. For example, rangefinding was properly conducted, sufficient anchor papers were provided, rater qualifications were demonstrated prior to scoring, raters were randomly presented with papers with consensus scores and scored them accurately. C. Requires a validity argument derived from the scoring rules, procedures, algorithm, consistency with human scoring, and/or approach to identifying and handling responses the algorithm is incapable of accurately scoring. Reference to current research/best practices is likely to be helpful. D. For example, consistency with human scoring. Often implemented through training based on human scored responses, human read-behind scoring, and/or human/machine hybrid scoring of responses. E. To demonstrate that student scores are reported in terms of the state's academic achievement standards. F. For all items other than those scored by humans or by machines emulating human scoring. Could include scoring rules for technology enhanced items, paper scanning procedures and criteria, quality control detection and follow-up for unclear responses on answer documents, special handling for soiled documents, etcetera.

Critical Element 4.5 – Multiple assessment forms

Critical Element	Examples of Evidence	Bite-sized Statements
If the State administers multiple forms[A] within a content area and grade level, within or across school years, the State ensures that all forms adequately represent the State's academic content standards[B] and yield consistent score interpretations such that the forms are comparable within[C] and across[D] school years.	Evidence to support this critical element for the State's assessments system includes documents such as: • Documentation of technically sound equating procedures and results within an academic year, such as a section of a technical report for the assessments that provides detailed technical information on the method used to establish linkages and on the accuracy of equating functions; [E] • As applicable, documentation of year-to-year equating procedures and results, such as a section of a technical report for the assessments that provides detailed technical information on the method used to establish linkages and on the accuracy of equating functions. [E]	**1. All forms for the same grade and subject are comparable within year** a. All forms within a year are aligned to the State content standards. b. Equating across forms within a year is sound. c. Technically sound planned horizontal equating practices were described. d. Equating procedures described were successfully carried out[K]. **2. All forms for the same grade and subject are comparable from year to year** a. Equating across forms across years is sound. b. Technically sound planned horizontal equating practices were described. c. Equating procedures described were successfully carried out[K].

Potential Sources of Evidence to Support Bite-sized Statements

4. Technical manuals, test specifications[F], test blueprints[G], alignment reports[J], equating procedures, equating evaluation reports[I], TAC minutes[H]
5. Technical manuals, test specifications, test blueprints, equating procedures, equating evaluation reports, TAC minutes

Notes

A. Depending on the state and vendor, the reasons for administering multiple forms can vary widely. For example, a state may have different forms for (1) time spans across the administration windows, (2) students who missed their original administration, (3) accommodations, (4) emergencies or breaches, (5) broadening content representation, or (6) the same time span to increase security. All forms, whatever the reason for them, should be included.
B. Looking for comparable alignment to the content standards across forms and from form to form. Some states have previously (and successfully) interpreted this to mean that across all forms (rather than on each form) the full breadth of the content standards is represented because the number of content standards was so large that it was necessary to identify a core set of standards that appear on all forms, and a remaining set of standards strategically assigned to only one form (a core and matrix approach). This argument relies on a sufficiently large core of standards appearing on all forms and representativeness from large content categories on each form to justify reasonable comparability across forms.
C. This asks for evidence of horizontal equating within years across forms.
D. This asks for evidence of horizontal equating of all forms across years.
E. Examples of documentation can 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 anchor items), the observed equating results (e.g., standards errors of equating across the achievement continuum, equating constants, properties of the linking/equating set), 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)
F. Test specifications clearly articulate what "equivalence" means from content (e.g., knowledge, skills and abilities), format and statistical perspectives.
G. Should contain documentation of the linkages from form to form (both within and across years).
H. May contain TAC deliberations on and/or endorsements of planned vertical and horizontal linkages, planned vertical and horizontal equating procedures, equating results, and/or results from independent validation/replication of equating.
I. Evaluation evidence may include the following: equating results from the vendor responsible for equating, independent review of results from the vendor responsible for equating, independent replication of results using the same software but with independently developed code, and independent replication of results using different software but the same methodology.
J. Will likely require that, at least in one grade, alignment studies demonstrate that alignment is achieved across forms.
K. Or, if a problem was encountered with the planned procedures, that the response successfully resolved the problem (along with independent validation of that approach).

Critical Element 4.6 – Multiple versions of an assessment

Critical Element	Examples of Evidence	Bite-sized Statements

If the State administers assessments in multiple versions[A] within a content area, grade level, or school year, the State:
• Followed a design and development process to support comparable interpretations of results for students tested across the versions of the assessments;
• Documented adequate evidence of comparability of the meaning and interpretations of the assessment results.
	Evidence to support this critical element for the State's general and alternate assessments includes:
	
	For the State's general assessments:
	• Documentation that the State followed a design and development process to support comparable interpretations of results across different versions of the assessments (e.g., technology-based and paper-based assessments, assessments in English and native language(s), general and alternate assessments based on grade-level academic achievement standards);
	  ○ For a native language assessment, this may include a description of the State's procedures for translation or trans-adaptation of the assessment or a report of analysis of results of back-translation of a translated test;
	  ○ For technology-based and paper-based assessments, this may include demonstration that the provision of paper-based substitutes for technology-enabled items elicits comparable response processes and produces an adequately aligned assessment;
	• Report of results of a comparability study of different versions of the assessments that is technically sound and documents evidence of comparability generally consistent with expectations of current professional standards.
	
	  If the State administers technology-based assessments that are delivered by different types of devices (e.g., desktop computers, laptops, tablets), evidence includes:
	• Documentation that test-administration hardware and software (e.g., screen resolution, interface, input devices) are standardized across unaccommodated administrations[B]; or
	• Either:
	  ○ Reports of research (quantitative or qualitative) that show that variations resulting from different types of delivery devices do not alter the interpretations of results; or
	  ○ A comparability study, as described above.
	
	AA-AAS. For the State's AA-AAS:
	Documentation that the State followed design, development and test administration procedures to ensure comparable results across different versions of the assessments, such as a description of the processes in the technical report for the assessments or a separate report[C].
		1. Item Development & Adaptation
		a. Stimulus[D] and item development training appropriately attune writers to avoiding anticipated issues in item adaptation.
		b. Stimulus and item development, submission, and acceptance procedures appropriately attend to anticipated item adaptation issues.
		c. Training was conducted and procedures were followed as described (or variability in implementation did not affect comparability).
		d. ***Computer-based assessment***: Training and procedures attended to anticipated item adaptation issues dealing with different hardware (e.g., computing, display, and input devices), and software (e.g., operating systems, interface, screen resolution).
		2. Empirical Studies of Item Adaptations
		a. Appropriate studies of student responses to adapted items were conducted[E].
		b. ***Computer-based assessment:*** student responses on different hardware and software variants are examined.
		c. Appropriate remedies were developed where issues were identified.
		3. Translation/Transadaptation
		a. Procedures for translation and/or transadaptation were developed specifically to attend to anticipated issues with translation and/or transadaptation.
		b. Procedures were followed as specified (or variability in implementation did not affect comparability).
		4. Test Design and Development
		a. Where blueprints could not be met using adapted items, replacement items were developed specifically for the adapted version to fulfill the same blueprint needs.
		5. Version Comparability[F]
		a. Technically sound studies of version comparability were designed and implemented.[G]
		b. Results demonstrate little to no practically significant measurement effects of adaptations (or appropriate adjustments were made).
		c. ***Computer-based assessment***: Comparability studies included hardware and software variants.

Critical Element 4.6 – Multiple versions of an assessment

Potential Sources of Evidence to Support Bite-sized Statements

1. Technical manuals, cognitive laboratory study design, item/stimulus development and review training materials, item/stimulus writing acceptance procedures and criteria, item/stimulus review procedures and criteria, audit trails for procedures and criteria, technical device requirement specifications, technical software requirements, technical reports, item pool sufficiency analyses.
2. Technical manuals, cognitive laboratory results, measurement invariance analysis procedures and results, DIF analysis procedures and results, technical reports.
3. Technical manuals, translation development, review, and certification procedures; transadaptation development, review, and certification procedures; independent spot-checks of translations and transadaptations. Includes Braille (as a pseudo language) and ASL (as a language).
4. Technical manuals, item pool sufficiency analyses, alignment analyses of adapted forms.
5. Technical manuals, measurement invariance analysis procedures/results, DIF and DTF (differential test functioning) analyses, equating studies (if adapted forms are not identical in content), other comparability analyses, and literature on comparability across hardware/software variations. ***For AA-AAS***: Studies of implementation fidelity.

Notes

A. Versions appear to include at the least the following variations in presentation:
   - Paper and pencil,
   - Computer-delivered,
   - Braille,
   - Translated (including ASL),
   - Transadapted (i.e., translated and adapted as necessary to make the version culturally appropriate),
   - Presentation on different devices (such as desktop, laptop, and tablets) and operating systems, and
   - Alternate assessments based on grade-level standards.

   For CAT, it is likely that only a subset of the total item pool will require adaptation. For all of these cases, states will need to present evidence that the various modes of presentation as well as new or adapted item pools and forms produce comparable results. In some cases this may require a logical argument based on procedure (because sample sizes are too small). For others, the number of variations may be too large to feasibly study all possibilities, and a strategic sampling framework may be useful to target variations most likely to be problematic.

B. This type of standardization is likely going too far. A more reasonable approach is to show that a set of minimum specifications were standardized (as explained in the second bullet).

C. This evidence example does not appear to add anything to what would be expected for general assessments. However, with AA-AAS assessments, sample sizes may be too small for traditional studies of comparability. See Note F.

D. Stimuli include such things as passages of text, tables, charts, graphs, figures, drawings, and photographs students refer to in responding to one or more test items.

E. Detailed studies of every item may not be feasible, but studies of strategically sampled items to address various item types, formats, and language load should be conducted. Studies may include such designs as cognitive laboratories or DIF studies where an adaptation is compared to the default mode. For very small populations (such as for Braille and ASL versions), expert judgment may instead be required. ***Computer-based assessment:*** Includes studies of hardware and software variations, and to maintain feasibility, states will likely need to make reasoned judgments about what constitutes an important variation and which variations can be grouped.

F. If items are adapted from a base form, then a combination of equating and measurement invariance studies could be used to provide evidence. For versions with large numbers of students (e.g., thousands), measurement invariance studies can include such studies as multigroup factor analyses. For versions with medium to large numbers of students (hundreds or more), they can include such studies as DIF or differential test functioning (DTF) studies. For versions with a small to large number (ten or more) of students they can include such studies as cognitive laboratories and expert judgment of equivalence of access to test content.

G. Unless states can prove that results of studies conducted by other entities are sufficiently applicable to the state's assessment, the evidence presented here should be studies completed using the empirical data from the current state assessment. This will be challenging depending on the number of possible modes of assessment that are delivered in your state. Additionally, states are expected to submit completed studies rather than planned studies. While qualitative data such as the results of cognitive laboratories may not be sufficient on its own, states should certainly submit this data if it is available.

Critical Element 4.7 – Technical analyses & ongoing maintenance

Critical Element	Examples of Evidence	Bite-sized Statements
The State has a system for monitoring and maintaining, and improving as needed, the quality of its assessment system, including clear and technically sound criteria for the analyses of all of the assessments in its assessment system (i.e., general assessments and alternate assessments).[A]	Evidence to support this critical element for the State's assessments system includes: • Documentation that the State has established and implemented clear and technically sound criteria for analyses of its assessment system, such as: o Sections from the State's assessment contract that specify the State's expectations for analyses to provide evidence of validity, reliability, and fairness; for independent studies of alignment and comparability, as appropriate; and for requirements for technical reports for the assessments and the content of such reports applicable to each administration of the assessment; o The most recent technical reports for the State's assessments that present technical analyses of the State's assessments; o Documentation of the alignment of the State's assessments to the State's academic content standards (e.g., evidence submitted under Critical Element 3.1 – Overall Validity, Including Validity Based on Content; o Presentations of assessments results (e.g., to the State's TAC); • Documentation of the State's system for monitoring and improving, as needed, the on-going quality of its assessment system, such as: o Evidence that the State has established and implemented clear criteria for the analysis of its assessment system (see above); o Documentation of regular internal and external technical review of components of the State's assessment system, such as State Board of Education minutes, minutes from TAC meetings, and documentation of roles and responsibilities of TAC members; o Outline of a deliberate cycle for reviewing and updating the State's academic content standards and assessments (e.g., provides for logical transitions such that the assessments are aligned to the standards on which instruction is based in the relevant school year).	**1. Current and Future Assessment System Quality** a. The state has criteria for acceptable results of current and future technical analyses of test quality.[B] b. The state has a documented system in place for reviewing technical quality and making improvements as needed (including to content standards).[B]

Potential Sources of Evidence to Support Bite-sized Statements

Internal policy documents, TAC bylaws or policies, TAC meeting minutes, technical reports, relevant qualifications and experience of the experts, contracts, meeting agendas for internal improvement reviews, documents showing content standards update review cycle.

Notes

A. This is asking states to set criteria for acceptable results of technical analyses of the quality of the state assessment system and to describe procedures put in place to improve state assessments when they do not meet the criteria. This goes beyond reporting on the results of technical analyses of assessment quality to mechanisms for deciding how good is good enough.

B. For example, clear rules are in place for evaluating the appropriateness of obtained reliability indices and estimates of precision – specifically regarding evidence of review by technical experts of these indices. Other examples would include alignment studies, standard setting results, equating results, and of course validity evidence.

Critical Element 5.1 – Procedures for including SWDs[A]

Critical Element	Examples of Evidence	Bite-sized Statements
The State has in place procedures to ensure the inclusion of all public elementary and secondary school students with disabilities in the State's assessment system, including, at a minimum, guidance for IEP Teams to inform decisions about student assessments that: • Provides clear explanations of the differences between assessments based on grade-level academic achievement standards and assessments based on alternate academic achievement standards, including any effects of State and local policies on a student's education resulting from taking an alternate assessment based on alternate academic achievement standards; • States that decisions about how to assess students with disabilities must be made by a student's IEP Team based on each student's individual needs; • Provides guidelines for determining whether to assess a student on the general assessment without accommodation(s), the general assessment with accommodation(s), or an alternate assessment; • Provides information on accessibility tools and features available to students in general and assessment accommodations available for students with disabilities; • Provides guidance regarding selection of appropriate accommodations for students with disabilities; • Includes instructions that students eligible to be assessed based on alternate academic achievement standards may be from any of the disability categories listed in the IDEA; • Ensures that parents of students with the most significant cognitive disabilities are informed that their student's achievement will be based on alternate academic achievement standards and of any possible consequences of taking the alternate assessments resulting from district or State policy (e.g., ineligibility for a regular high school diploma if the student does not demonstrate proficiency in the content area on the State's general assessments); • The State has procedures in place to ensure that its implementation of alternate academic achievement standards for students with the most significant cognitive disabilities promotes student access to the general curriculum.[B]	Evidence to support this critical element for the State's assessment system includes: • Documentation that the State has in place procedures to ensure the inclusion of all students with disabilities, such as: ○ Guidance for IEP Teams and IEP templates for students in tested grades; ○ Training materials for IEP Teams; ○ Accommodations manuals or other key documents that provide information on accommodations for students with disabilities; ○ Test administration manuals or other key documents that provide information on available accessibility tools and features; • Documentation that the implementation of the State's alternate academic achievement standards[C] promotes student access to the general curriculum, such as: ○ State policies that require that instruction for students with the most significant cognitive disabilities be linked to the State's grade-level academic content standards; ○ State policies that require standards-based IEPs linked to the State's grade-level academic content standards for students with the most significant cognitive disabilities; ○ Reports of State monitoring of IEPs that document the implementation of IEPs linked to the State's grade-level academic content standards for students with the most significant cognitive disabilities. Note: Key topics related to the assessment of students with disabilities are also addressed in Critical Element 4.2 -- Fairness and Accessibility and in critical elements addressing the AA-AAS throughout.	**1. Inclusion of all SWDs** a. All public elementary and secondary students are required to be included in the state assessment system.[D] **2. Guidance to IEP teams on assessment selection** a. Provides guidance for selecting between general assessment without accommodations, general assessment with accommodations, or alternate assessment for any student with a disability listed in IDEA.[E] b. Provides guidance on the selection of accommodations.[E,F] c. Explains differences between grade-level and alternate assessments. d. Explains available accessibility tools and accommodations.[G] e. Clearly explains effects of taking alternate assessments to both educators and parents (e.g., on educational programming and/or eligibility for diploma). f. Leaves the general/alternate and accommodations/no accommodations decisions to the IEP team.[E] **3. Policies promote access to the general curriculum for the AA-AAS population**[H] a. Instruction of students in the AA-AAS population is required to be linked to grade level content standards. b. IEPs for the AA-AAS population are required to be standards-based and to be linked to grade level content standards. c. Documentation of state processes and procedures ensure implementation of state policies.

Critical Element 5.1 – Procedures for including SWDs[A]

Potential Sources of Evidence to Support Bite-sized Statements
1. Binding legislation, regulation, rule, or policy; accessibility and accommodations manuals; IEP requirements documents, templates, training materials, test administration manuals, and guidance 2. IEP requirements documents, templates, training, and guidance. 3. Binding legislation, regulation, rule, or policy; accessibility and/or accommodations manuals.

Notes
A. Nothing in this guidance appears to preclude the use of portfolio-based assessments for the AA-AAS so long as the portfolio-based assessments can meet the peer review requirements. B. This bullet is qualitatively different than the rest of the bullets addressing guidance to IEP teams (it should probably be treated as a separate paragraph). C. For states where parental refusal (opt-out) is allowable, states may need to explain attempts to balance the preservation of inclusion policies and parental rights. ED has indicated that the 95% participation requirement remains, and that two avenues of action may be taken: (1) review by peers of the adequacy of state guidance on participation, and (2) review of state compliance outside of peer review. However, ED has also indicated that the effect of having less than 95% participation is not yet clear. D. These three statements (i.e., A., B., and F.) may appear to be in conflict. However, together they mean the guidance to the IEP team (e.g., suggested flowcharts, suggested criteria for decision-making) must be clear, but decisions about the appropriate assessment and/or appropriate accommodations based on that guidance legally rest with the IEP team. E. Clear and standardized guidelines are in place to support the assignment of students to appropriate accommodations. Guidelines should clearly indicate conditions under which particular accommodations should be used. Guidelines should be in accordance with all federal and state laws and polices concerning the access to and assignment of accommodations. Tools should be provided within the guidelines to readily available to assist with decision-making (e.g., decision-making trees, FAQs). F. May come from binding policy documents such as statute, rule, executive order, IEP requirements, and/or IEP implementation requirements. G. Evidence of implementation is likely to require, at a minimum, monitoring of IEPs through a random sampling approach (possibly alongside examinations of IEP implementation fidelity). State's offices of special education should have monitoring protocols, and possibly monitoring reports, that could be submitted as evidence. H. If states use the same AA-AAS form for multiple grades in a grade band rather than a separate assessment for each grade, a rationale and evidence for why this is appropriate and why it promotes access to the general curriculum should be provided.

Critical Element 5.2 – Procedures for including ELs

Critical Element	Examples of Evidence	Bite-sized Statements
The State has in place procedures to ensure the inclusion of all[A] English learners in public elementary and secondary schools in the State's assessment system and clearly communicates this information to districts, schools, teachers, and parents, including, at a minimum: • Procedures for determining whether an English learner should be assessed with accommodation(s); • Information on accessibility tools and features available to all students and assessment accommodations available for English learners; • Guidance regarding selection of appropriate accommodations for English learners.	Evidence to support this critical element for the State's assessment system includes: • Documentation of procedures for determining student eligibility for accommodations and guidance on selection of appropriate accommodations for English learners; • Accommodations manuals or other key documents that provide information on accommodations for English learners; • Test administration manuals or other key documents that provide information on available accessibility tools and features; • Guidance in key documents that indicates all accommodation decisions must be based on individual student needs and provides suggestions regarding what types of accommodations may be most appropriate for students with various levels of proficiency in their first language and English. Note: Key topics related to the assessment of English learners are also addressed in Critical Element 4.2 – Fairness and Accessibility.	**1. Inclusion of all ELs** a. All[A] public elementary and secondary EL students are required to be included in the state assessment system[B]. b. Requirements and procedures are clearly communicated to districts, schools, teachers, and parents. **2. Guidance on accommodations** a. Documentation and procedures provide guidance for determining eligibility for accommodations and selecting appropriate accommodations (preferably with suggestions based on English proficiency level and/or native language proficiency level). b. Documentation and procedures explains available accommodations[C]. c. Documentation and procedures indicates that final decisions on accommodations must be based on individual student needs. **3. Guidance on Accessibility Tools** a. Documentation and procedures explain available accessibility tools for all students if applicable to ELs.

Potential Sources of Evidence to Support Bite-sized Statements

1. Binding legislation, regulation, rule, or policy; accommodations manuals; test administration manuals; official communications to schools, districts, and parents; participation data for ELLs.
2. Assessment inclusion policies, accommodations manuals, guidance/assistance for selecting accommodations, documentation of translated communications for parents.
3. Accommodations manuals, official communication of available accessibility tools, guidance for using accessibility tools.

Notes

A. This peer review guidance element does not address the reading/ELA exception for ELs who were eligible for, and took, an English Language Proficiency (ELP) test (i.e., first-year immigrant ELs or ELs with a case-by-case extension). Clarification from USDE has confirmed that peer review does not currently include a review of ELP assessments, though it is something the department has signaled it is interested in considering in the future.

B. For states where parental refusal (opt-out) is allowable, states may need to explain attempts to balance the preservation of inclusion policies and parental rights. ED has indicated that the 95% participation requirement remains, and that two avenues of action may be taken: (1) review by peers of the adequacy of state guidance on participation, and (2) review of state compliance outside of peer review. However, ED has also indicated that the effect of having less than 95% participation

C. For states administering ACT, it may be appropriate to communicate to all stakeholders that accommodations for EL students will result in scores that cannot be used for college admissions.

Critical Element 5.3 – Accommodations

Critical Element	Examples of Evidence	Bite-sized Statements
The State makes available appropriate accommodations and ensures that its assessments are accessible to students with disabilities and English learners. Specifically, the State: • Ensures that appropriate accommodations are available for students with disabilities under IDEA and students covered by Section 504; • Ensures that appropriate accommodations are available for English learners; • Has determined that the accommodations it provides (i) are appropriate and effective for meeting the individual student's need(s) to participate in the assessments, (ii) do not alter the construct being assessed, and (iii) allow meaningful interpretations of results and comparison of scores for students who need and receive accommodations and students who do not need and do not receive accommodations[A]; • Has a process to individually review and allow exceptional requests for a small number of students who require accommodations beyond those routinely allowed.	Evidence to support this critical element for both the State's general and AA-AAS includes: • Lists of accommodations available for students with disabilities under IDEA, students covered by Section 504 and English learners that are appropriate and effective for addressing barrier(s) faced by individual students (i.e., disability and/or language barriers) and appropriate for the assessment mode (e.g., paper-based vs. technology-based), such as lists of types of available accommodations in an accommodations manual, test coordinators manual or test administrators manual; • Documentation that scores for students based on assessments administered with allowable accommodations (and accessibility tools and features, as applicable) allow for valid inferences, such as: ○ Description of the reasonable and appropriate basis[A] for the set of accommodations offered on the assessments, such as a literature review, empirical research, recommendations by advocacy and professional organizations, and/or consultations with the State's TAC, as documented in a section on test design and development in the technical report for the assessments; ○ For accommodations not commonly used in large-scale State assessments, not commonly used in the manner adopted for the State's assessment system, or newly developed accommodations, reports of studies, data analyses, or other evidence[A] that indicate that scores based on accommodated and non-accommodated administrations can be meaningfully compared; ○ A summary of the frequency of use of each accommodation on the State's assessments by student characteristics (e.g., students with disabilities, English learners); • Evidence that the State has a process to review and approve requests for assessment accommodations beyond those routinely allowed, such as documentation of the State's process as communicated to district and school test coordinators and test administrators[B].	**1. Accommodations are provided as appropriate** c. For SWDs covered under IDEA. d. For SWDs covered under Section 504. e. For ELs[C] **2. Determinations of appropriateness** a. Accommodations are appropriate and effective in reducing barriers to access for the intended populations[D] b. Accommodations do not alter the intended construct.[E] c. Accommodated tests produce results comparable to those for students who appropriately take the test *without* accommodations.[E] **3. Exceptional accommodations requests** a. The state has an appropriate process for reviewing and ruling on requests for individual students to receive additional accommodations beyond those typically allowed. b. The availability and details of the process are clearly communicated to district and school stakeholders.

Potential Sources of Evidence to Support Bite-sized Statements

1. Accommodations manual; technical manual; accommodations selection guidance; IEP/504/EL requirements, templates, training materials; official communications to districts and schools, FAQs.
2. Technical manual, measurement invariance studies[E], literature reviews[E], cognitive laboratories[E], expert/advocate judgment[E], studies of differential boost[E], guidelines for interpreters.
3. Official communications to district and school testing coordinators, exceptional accommodations request process and approval criteria, audit trails for compliance with process and criteria, guidelines for interpreters.

Notes

A. Validation that individual accommodations are appropriate and effective in reducing barriers to access, and that their use does not modify the intended construct is notoriously difficult. This is especially so when the number of students using a given accommodation is small. The wording used by USED (*has determined*, *reasonable and appropriate basis*, *other evidence*) appears to leave some practical flexibility for these criteria to be evaluated at varying levels of rigor, depending on the practicality of the various methods

Critical Element 5.3 – Accommodations

Notes
of evaluation. Evaluations might occur empirically (e.g., studies of differential, boost, studies of measurement invariance such as DIF and multi-group confirmatory factor analyses, or literature review), qualitatively (e.g., cognitive laboratories, literature review), or subjectively but not arbitrarily (e.g., expert opinion from professional organizations and/or advocacy groups). More rigorous methods of determining appropriateness increase the strength of the submission. Evidence may include documentation that scores assigned to students representing subgroups of the population have similar levels of reliability and validity when compared to those associated with that total population. B. Should include methods for subjectively, but not arbitrarily, determining that the requested exceptional accommodation meets the three criteria applied to regularly available accommodations: *i. Is appropriate and effective for meeting the individual student's need(s) to participate in the assessments and in classroom work (i.e., reduces barriers to access).* *ii. Does not alter the construct being assessed.* *iii. Allows meaningful interpretations of results and comparison of scores for students who need and receive accommodations and students who do not need and do not receive accommodations.* Could also include documentation of that non-approved accommodations are not considered valid, and state procedures for handling such scores. C. 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. D. Evidence is collected showing that each accommodation provides the support intended. Evidence suggests that the accommodation provides for gains in the intended populations. E. Different types of studies are appropriate for accommodations with different sample sizes. For moderate to large sample sizes (e.g., hundreds or more) studies of differential boost and measurement invariance may be appropriate. For small to large sample sizes (sample sizes of one and above), subjective methods such as cognitive laboratories (demonstrating increased access for SWDs and ELs) and/or expert or advocate judgment may be more appropriate. In all cases, literature reviews are appropriate. Because a large number of accommodations may be available, for feasibility, states may need to prioritize (and justify the prioritization) of accommodations to be evaluated.

Critical Element 5.4 – Monitoring Test Administration for Special Populations[A]

Critical Element	Examples of Evidence	Bite-sized Statements
The State monitors test administration in its districts and schools to ensure that appropriate assessments, with or without appropriate accommodations, are selected for students with disabilities under IDEA, students covered by Section 504, and English learners so that they are appropriately included in assessments and receive accommodations that are[B]: • Consistent with the State's policies for accommodations; • Appropriate for addressing a student's disability or language needs for each assessment administered; • Consistent with accommodations provided to the students during instruction and/or practice; • Consistent with the assessment accommodations identified by a student's IEP Team or 504 team for students with disabilities, or another process for an English learner; • Administered with fidelity to test administration procedures.	Evidence to support this critical element for the State's assessment system includes documents such as[B]: • Description of procedures the State uses to monitor that accommodations selected for students with disabilities, students covered by Section 504, and English learners are appropriate; • Description of procedures the State uses to monitor that students with disabilities are placed by IEP Teams in the appropriate assessment; • The State's written procedures for monitoring the use of accommodations during test administration, such as guidance provided to districts; instructions and protocols for State, district and school staff; and schedules for monitoring; • Summary of results of monitoring for the most recent year of test administration in the State.	**1. Appropriate selection of assessment and/or accommodations**[B] a. IEP and 504 teams follow appropriate procedures to select an appropriate assessment. b. IEP, 504, and EL teams follow appropriate procedures to designate accommodations. c. Designated accommodations are consistent with those received as part of instruction.[C] **2. Appropriate provision of general/alternate assessments and/or accommodations**[D,E] a. SWDs are included in assessment. b. ELs are included in assessment. c. SWDs receive the IEP-designated assessment (general/alternate). d. SWDs receive the IEP-designated or 504 plan-designated accommodations. e. ELs receive the designated accommodations. f. Accommodations are administered in accordance with approved procedures.

Potential Sources of Evidence to Support Bite-sized Statements

1. Technical reports, IEP/504/EL plan monitoring procedures and results, EL accommodations selection monitoring procedures, sampling plans for monitoring, TAC minutes, accommodations manuals, and LEA policy documents.
2. Technical reports, assessment administration monitoring procedures and results, sampling plans for monitoring, TAC minutes, directions for test interpreters, summaries of test taking times (e.g., within school hours) and amount of time taking the test by SWD/non-SWD and by EL/non-EL, summaries of monitoring results, summaries of irregularity reports, and summaries of appeals.

Notes

A. This critical element will be checked by ED and not go to the team of peers unless the evidence provided needs additional review.
B. Because of the number of SWDs with IEPs and 504 plans and the number of ELs, all statements in this section will likely need to be supported through sparse sampling. This may include reviewing a sample of IEPs, 504 plans, and EL plans, instructional accommodations provided for those students, and observing the provision of test accommodations to those students.
C. Guidelines are in accordance with all state laws and policies concerning the access to and assignment of accommodations which typically require an alignment across IEP determined instructional accommodations and IEP determined assessment accommodations.
D. When an accommodation is presented in the form of an accessibility feature offered to all students (e.g., read-aloud), procedures are in place to ensure that students who require that accommodation are familiar with how and when to use and access that accessibility feature.
E. In order to demonstrate evidence related to this critical element, the state will likely need to have procedures in place for associating student identifiers with the types of accommodation(s) or accessibility features that were used for the test in post-test validation studies.

Critical Element 6.1 – State Adoption of Academic Achievement Standards for All Students

Critical Element	Examples of Evidence	Bite-sized Statements
The State formally adopted challenging academic achievement standards in reading/language arts, mathematics and in science[A] for all students, specifically: • The State formally adopted academic achievement standards in the required tested grades and, at its option, also alternate academic achievement standards for students with the most significant cognitive disabilities; • The State applies its grade-level academic achievement standards to all public elementary and secondary school students enrolled in the grade to which they apply, with the exception of students with the most significant cognitive disabilities to whom alternate academic achievement standards may apply; • The State's academic achievement standards and, as applicable, alternate academic achievement standards, include: (a) At least three levels of achievement, with two for high achievement and a third for lower achievement; (b) descriptions of the competencies associated with each achievement level; and (c) achievement scores that differentiate among the achievement levels.	Evidence to support this critical element for the State's assessment system includes: • Evidence of adoption of the State's academic achievement standards and, as applicable, alternate academic achievement standards, in the required tested grades and subjects (i.e., in reading/language arts and mathematics for each of grades 3-8 and high school and in science for each of three grade spans (3-5, 6-9, and 10-12)), such as State Board of Education minutes, memo announcing formal approval from the Chief State School Officer to districts, legislation, regulations, or other binding approval of academic achievement standards and, as applicable, alternate academic achievement standards; • State statutes, regulations, policy memos, State Board of Education minutes, memo from the Chief State School Officer to districts or other key documents that clearly state that the State's academic achievement standards apply to all public elementary and secondary school students in the State (with the exception of students with the most significant cognitive disabilities to whom alternate academic achievement standards may apply); • Evidence regarding the academic achievement standards and, as applicable, alternate academic achievement standards, regarding: (a) at least three levels of achievement, including two levels of high achievement (e.g., proficient and advanced) and a third of lower achievement (e.g., basic); (b) descriptions of the competencies associated with each achievement level; and (c) achievement scores (i.e., "cut scores") that differentiate among the achievement levels.	**1. Adoption of Academic Achievement Standards** a. Academic achievement standards have been formally adopted in reading/language arts, mathematics, and science in all required tested grades. b. Academic achievement standards apply to all students (through either general or alternate assessments). **2. Characteristics of Academic Achievement Standards** a. Academic achievement standards are challenging. See Critical Element 6.3. b. Academic achievement standards delineate at least three achievement levels (with at least two indicating proficient achievement, and at least one indicating non-proficient achievement). c. Each achievement level for each grade and subject is associated with an achievement level descriptor. d. Achievement levels are differentiated by cut scores.

Potential Sources of Evidence to Support Bite-sized Statements

1. State board of education minutes; binding legislation, regulation, rule, or policy; public memo announcing adoption; public document describing to whom standards apply.
2. Achievement level descriptors[B], cut score tables, standard setting report, percentages in performance levels, changes in percentages in performance levels from previous cut scores, comparison of percentages in performance levels and/or cut scores to external benchmarks (e.g., NAEP reports, college readiness reports, NAEP cut scores, college readiness benchmarks, AP/IB scores, college grades).

Notes

A. Different documentation may be needed for each content area depending upon whether the standards in each content area have changed since peer review and/or ESEA flexibility approval. For example, a state may have updated its reading/language arts and mathematics standards since receiving approval, but not its science standards. In this case, a peer review or ESEA flexibility approval letter would suffice for science, but new evidence would be required for ELA and mathematics.

B. Sometimes also called "Performance Level Descriptors (PLDs)." Score reports with links to websites of extended descriptors, are not sufficient. In this case, the text of the extended descriptors should be submitted directly.

Critical Element 6.2 – Achievement standards setting

Critical Element	Examples of Evidence	Bite-sized Statements
The State used a technically sound method and process that involved panelists with appropriate experience and expertise for setting its academic achievement standards and alternate academic achievement standards to ensure they are valid and reliable.	Evidence to support this critical element for the State's general assessments and AA-AAS includes: • The State's standards-setting report, including: ○ A description of the standards-setting method and process used by the State; ○ The rationale for the method selected; ○ Documentation that the method used for setting cut scores allowed panelists to apply their knowledge and experience in a reasonable manner and supported the establishment of reasonable and defensible cut scores; ○ Documentation of the process used for setting cut scores and developing performance-level descriptors aligned to the State's academic content standards; ○ A description of the process for selecting panelists; ○ Documentation that the standards-setting panels consisted of panelists with appropriate experience and expertise, including: ▪ Content experts with experience teaching the State's academic content standards in the tested grades; ▪ Individuals with experience and expertise teaching students with disabilities, English learners and other student populations in the State; ▪ As appropriate, individuals from institutions of higher education (IHE) and individuals knowledgeable about career-readiness; ▪ A description, by relevant characteristics, of the panelists (overall and by individual panels) who participated in achievement standards setting; ○ If available, a summary of statistical descriptions and analyses that provides evidence of the reliability[A] of the cut scores and the validity of recommended interpretations.[B] **AA-AAS.** For the State's AA-AAS, in addition to the above: • Documentation that the panels for setting alternate academic achievement standards included individuals knowledgeable about the State's academic content standards and special educators knowledgeable about students with the most significant cognitive disabilities.	**1. Standard setting process** a. The state used a technically sound and well-documented process to develop reasonable, defensible i. achievement-level descriptors aligned to the content standards, and ii. achievement standards (i.e., cut scores).[C] b. The cut scores are adequately reliable.[A] c. The cut scores adequately distinguish between the achievement levels as described by the achievement level descriptors.[B] **2. Standard setting participants** a. Participants, across and within panels, had the appropriate experience and expertise to engage appropriately in the process.[D] b. Participants were able to appropriately apply their knowledge and skills to the tasks.[E]

Potential Sources of Evidence to Support Bite-sized Statements

1. Standard setting report, including standard setting process/script, standard setting training materials/script, participant rating summaries; achievement level descriptor development process/script TAC minutes, item mapping studies[F].
2. Descriptive summaries of achievement level standards setting panelists with relevant characteristics, rater evaluations of the process and their use of their expertise, for each panel and for the totality of panels.

Notes

A. For example, that standard errors around recommended cut scores were reasonably small (and, possibly, that they converged from one round to the next).
B. The validity of recommended interpretations are based on studies that examine the claims associated with cut scores (e.g, that a given cut score indicates students are academically college and career ready or that descriptors describe evidence elicited by items). The validity of interpretations is detailed under critical element 3.1.
C. The required evidence on the development of the achievement-level descriptors and achievement standards is likely to be substantial. ED has not specified what this evidence should look like in great detail, but consistency with all aspects of the *Standards* (see p. 95–107) applicable to standard setting will likely be expected by peer reviewers.
D. Important considerations include the various roles of stakeholders that will use test scores, adequate content knowledge, representation from educators of students from various demographic groups, and representation of panelists from various demographic groups.
E. Could be supported using summaries of participant self-ratings or a report of an independent auditor.
F. Studies documenting the types of knowledge and cognitive complexity represented by items with difficulties at different locations across the score scale.

Critical Element 6.3 – Challenging & aligned academic achievement standards[A]

Critical Element	Examples of Evidence	Bite-sized Statements
The State's academic achievement standards are challenging and aligned with the State's academic content standards such that a high school student who scores at the proficient or above level has mastered what students are expected to know and be able to do by the time they graduate from high school in order to succeed in college and the workforce[B]. If the State has defined alternate academic achievement standards for students with the most significant cognitive disabilities, the alternate academic achievement standards are linked to the State's grade-level academic content standards or extended academic content standards, show linkage to different content across grades, and reflect professional judgment of the highest achievement standards possible for students with the most significant cognitive disabilities[C].	Evidence to support this critical element for the State's general assessments and AA-AAS includes: For the State's general assessments: • Documentation that the State's academic achievement standards are aligned with the State's academic content standards, such as: ○ A description of the process used to develop the State's academic achievement standards that shows that: ▪ The State's grade-level academic content standards were used as a main reference in writing performance level descriptors; ▪ The process of setting cut scores used, as a main reference, performance level descriptors that reflect the State's grade-level academic content standards; ▪ The State's cut scores were set and performance level descriptors written to reflect the full range of the State's academic content standards for each grade[D]; ○ A description of steps taken to vertically articulate the performance level descriptors across grades; ○ Evaluation by standard-setting panelists or external expert reviewers that the State's academic achievement standards are aligned to the grade-level academic content standards and include subject-specific performance level descriptors that meaningfully differentiate across performance levels within grades and are vertically articulated across grades; • Documentation that the State's academic achievement standards are challenging[E], such as: ○ Reports of the results of benchmarking the State's academic achievement standards against NAEP, international assessments or other related and appropriate measures; ○ Policies of the State network of institutions of higher education (IHEs) that exempt from remedial courses and place into credit-bearing college courses any student who scores at the proficient level or above on the State's high school assessments. **AA-AAS.** For the State's AA-AAS[F]: • Documentation that the State's alternate academic achievement standards are linked to the State's academic content standards, such as: ○ A description of the process used to develop the alternate academic achievement standards that shows: ▪ The State's grade-level academic content standards or extended academic content standards were used as a main reference in writing performance level descriptors for the alternate academic achievement standards; ▪ The process of setting cut scores used, as a main reference, performance level descriptors linked to the State's grade-level academic content standards or extended academic content standards; ▪ The cut scores were set and performance level descriptors written to link to the State's grade-level academic content standards or extended academic content standards; ▪ A description of steps taken to vertically articulate the alternate academic achievement standards (including cut scores and performance level descriptors) across grades.	**1. Challenging Achievement Standards** a. Performance or Achievement level descriptors (ALDs) fully represent the knowledge and skills in the content standards.[G] b. Achievement standards (cut scores) are aligned to the content standards.[E] c. **AA-AAS**: ALDs represent the highest levels of achievement appropriate for the AA-AAS population.[I] **2. Meaningful Achievement Standards** a. ALDs are vertically articulated across grades.[G] b. Achievement standards (cut scores) meaningfully differentiate between levels of achievement within grade level. c. "Proficient" and above ALDs in high school represent the academic knowledge and skills necessary to succeed in college and the workforce. d. Achievement standards (cut scores) are vertically articulated across grade levels.[H] **3. Aligned Achievement Standards** a. ALDs are clearly derived from, and represent the full range of, State content standards. b. The standard setting process considered the full range of State content standards.

Critical Element 6.3 – Challenging & aligned academic achievement standards[A]

Potential Sources of Evidence to Support Bite-sized Statements
1. ALDs, ALD development process, ALD development report, comparison to ALDs for programs considered rigorous (e.g., NAEP, college readiness benchmarks, PISA, AP, IB); standard setting report that describes the identification of high school "proficient" cut scores that align with college/workforce readiness through the use of external benchmarks such as SAT or ACT, role of post-secondary educators in recommending high school cut scores, or other such process; expert opinion for AA-AAS. 2. Vertical articulation process for ALD development, vertical articulation process in standard setting, analyses of separability of cut scores, item mapping studies comparing performance levels[J], percentages in performance levels, analyses comparing cut scores to external benchmarks considered rigorous (e.g., NAEP, PISA, AP, IB, college readiness benchmarks, college grades), TAC minutes. 3. ALD development process, crosswalks between ALDs and content standards/content specifications, alignment analyses (to show achievement standards spanned the full range of content standards).

Notes
A. This is an element on which ED did not include the word "collectively" to indicate that a validity argument including all evidence is needed. However, it is likely that the peer reviewers will be looking for more than just each Bite-sized Statement being supported. It is likely that the peers will be looking for a synthesis of all evidence in this critical element to support the validity of cut scores. B. May require endorsement by institutions of higher education (IHEs) for high school achievement level descriptors and achievement standards (whether ED has authority to require this for states without an ESEA waiver is unclear) or a description of how standards were set to ensure the state's target for achievement standards was judged to be appropriate and defensible. C. The wording of this critical element suggests that professional judgment that the cut scores on the AA-AAS represent the highest achievement standards possible for the AA-AAS population is sufficient. D. Writing should be assessed in all grades in which writing is addressed in the State's standards. If listening and speaking are addressed in the standards, but not assessed, States will need to include a discussion and plan of how those skills will be assessed in the future (see element 2.1). E. Could be supported by analyses of cognitive complexity, impact data, benchmarking to external national/international criteria considered rigorous (i.e., NAEP proficiency, PISA benchmarks, AP performance levels, ACT college readiness benchmarks, SAT college benchmarks), and/or endorsement by IHEs. F. This does not appear to require anything in addition to that required for general assessments. G. For example, the "Proficient" achievement level descriptors in each grade could be designed following a progression that ends at the knowledge and skills needed for success in college and the workforce. H. Concerns here would be lower performance in a lower grade relative to a higher grade (e.g., mean reversals on a vertical scale), considerable differences in impact data from one grade to the next without a sound rationale, and lack of consistent classifications from one grade to the next and from high school to post-secondary outcomes. I. If the state has defined alternate achievement standards for students with the most significant cognitive disabilities, the success of ALDs for AA-AAS hinges on the adequacy those alternate achievement standards in represent the highest level of knowledge and skills appropriate for the AA-AAS population. J. To demonstrate that the differences between performance levels are quantitatively meaningful.

Critical Element 6.4 - Reporting

Critical Element	Examples of Evidence	Bite-sized Statements
The State reports its assessment results, and the reporting facilitates timely, appropriate, credible, and defensible interpretations and uses of results for students tested by parents, educators, State officials, policymakers and other stakeholders, and the public, including: • The State reports to the public its assessment results on student achievement at each proficiency level and the percentage of students not tested for all students and each student group after each test administration; • The State reports assessment results, including itemized score analyses[B], to districts and schools so that parents, teachers, principals, and administrators can interpret the results and address the specific academic needs of students, and the State also provides interpretive guides to support appropriate uses of the assessment results; • The State provides for the production and delivery of individual student interpretive, descriptive, and diagnostic reports after each administration of its assessments that: ○ Provide valid and reliable information regarding a student's achievement; ○ Report the student's achievement in terms of the State's grade-level academic achievement standards (including performance-level descriptors); ○ Provide information to help	**Collectively**[A], for the State's assessment system, evidence to support this critical element must demonstrate that the State's reporting system facilitates timely, appropriate, credible, and defensible interpretation and use of its assessment results. Evidence to support this critical element both the State's general assessments and AA-AAS includes: • Evidence that the State reports to the public its assessment results on student achievement at each proficiency level and the percentage of students not tested for all students and each student group after each test administration, such as: ○ State report(s) of assessment results; ○ Appropriate interpretive guidance provided in or with the State report(s) that addresses appropriate uses and limitations of the data (e.g., when comparisons across student groups of different sizes are and are not appropriate). • Evidence that the State reports results for use in instruction, such as: ○ Instructions for districts, schools, and teachers for access to assessment results, such as an electronic database of results; ○ Examples of reports of assessment results at the classroom, school, district and State levels provided to teachers, principals, and administrators that include itemized score analyses[B], results according to proficiency levels, performance level descriptors, and, as appropriate, other analyses that go beyond the total score (e.g., analysis of results by strand); ○ Instructions for teachers, principals and administrators on the appropriate interpretations and uses of results for students tested that include: the purpose and content of the assessments; guidance for interpreting the results; appropriate uses and limitations of the data; and information to allow use of the assessment results appropriately for addressing the specific academic needs of students, student groups, schools and districts. ○ Timeline that shows results are reported to districts, schools, and teachers in time to allow for the use of the results in planning for the following school year. • Evidence to support this critical element for both general assessments and AA-AAS, such as: ○ Templates or sample individual student reports for each content area and grade level for reporting student performance that: ▪ Report on student achievement according to the domains and subdomains defined in the State's academic content standards and the achievement levels for the student scores (though sub-scores should only be reported when they are based on a sufficient number of items or score points to provide valid and reliable results); ▪ Report on the student's achievement in terms of grade-level achievement using the State's grade-level academic achievement standards and corresponding performance level descriptors; ▪ Display information in a uniform format and use simple language that is free of jargon and understandable to parents, teachers, and principals; ▪ Examples of the interpretive guidance that accompanies individual student reports, either integrated with the report or a separate page(s), including cautions related to the	**1. Reporting Logistics** a. The state publicly reports its assessment results. b. Reports include summaries of students at each proficiency level. c. Reports include summaries of students not tested for all students and each subgroup. d. The state follows a defined process to ensure that individual student reports are provided to schools, teachers, and parents in a timely manner after every administration.[C] **2. Reporting Supports Timely[D], Appropriate, Credible, and Defensible Interpretation and Use of Assessment Data**[A] a. Reports are useful for relevant stakeholders for intended purposes.[E] b. Itemized reports[B] and data made available to educators are appropriate to inform instructional/educational decisions for students.[E] c. Reports are accompanied by interpretive guidance regarding purpose and content, meaning of scores, appropriate interpretations and uses, and (potentially) common inappropriate interpretations and uses.[F] **3. Individual Student Reports** a. Individual student reports are interpretive, descriptive, and diagnostic.[G,B] b. Individual student reports provide valid and reliable information about student achievement. c. Are based on the appropriate content standards (i.e., general or

Critical Element 6.4 - Reporting

Critical Element	Examples of Evidence	Bite-sized Statements
parents, teachers, and principals interpret the test results and address the specific academic needs of students; ○ Are available in alternate formats (e.g., Braille or large print) upon request and, to the extent practicable, in a native language that parents can understand; ○ The State follows a process and timeline for delivering individual student reports to parents, teachers, and principals as soon as practicable after each test administration.	reliability of the reported scores; ▪ Samples of individual student reports in other languages and/or in alternative formats, as applicable. • Evidence that the State follows a process and timeline for delivering individual student reports, such as: ○ Timeline adhering to the need for the prompt release of assessment results that shows when individual student reports are delivered to districts and schools; ○ Key documents, such as a cover memo that accompanies individual student reports delivered to districts and schools, listserv messages to district and school test coordinators, or other meaningful communication to districts and schools that include the expectation that individual student reports be delivered to teachers and principals and corresponding expectations for timely delivery to parents (e.g., within 30 days of receipt). Note: Samples of individual student reports and any other sample reports should be redacted to protect personally identifiable information, as appropriate, or populated with information about a fictitious student for illustrative purposes.	alternate) for the individual student d. Include achievement level descriptors. e. Provide interpretive guidance to parents, teachers, and principals. f. Are useful to address specific academic needs of students and groups of students (e.g., classrooms, subgroups, schools, districts).[B,E,F,G] g. Are made available in alternate formats when requested. h. Are made available in a parent's native language, where practicable.

Potential Sources of Evidence to Support Bite-sized Statements

1. Sample publicly accessible reports, interpretive guides, press releases, policies for distribution of individual reports, official communications to schools and districts.
2. Samples reports (mocked up or with PII redacted); interpretive guides; official communications to districts, schools, and parents; instructions for retrieving reports/data files; report development processes, criteria, and stakeholder involvement[E], assistance provided for analyzing data files[H], documentation of locally developed tools and/or databases using test data for educational/instructional planning.
3. Sample individual student/parent reports (mocked up or with PII redacted), summary of focus groups' feedback on report usability and design, instructions for educators to retrieve reports, interpretive guides, assistance provided for analyzing data files[H].

Notes

A. The entire set of evidence around reporting should be synthesized to build an argument that the reporting system (reports/data files) meets this requirement overall.
B. "Itemized score analyses" does not necessarily mean item-level reports, but could include item maps (descriptors of content items measure near specific test scores), subscore (e.g., strand) reporting, analysis of strengths/weaknesses, or other reports beyond overall score reporting. It is legitimate to argue against item level or subscore reporting on the bases of psychometric analyses of score stability (reliability), meaningfulness (validity), and/or deleterious impacts of using unreliable/ invalid data in consequential decisions. Score reports should support inferences relative to key content standards by reporting at the finest grain size supported by reliability and validity evidence. Audience-appropriate reliability/precision information is provided for each reported score (including sub-scores and growth scores) to facilitate intended interpretations.
C. Including broad distribution of clear directions for accessing and viewing score reports.
D. 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. A first-year delay in reporting to allow for standard setting can be defended as a need to ensure results can be interpreted and used as intended.
E. For example, feedback had been gathered from a 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 reports to answer questions).
F. Inappropriate uses should be cautioned against, and documentation provided of the cautions provided.
G. These should be interpreted broadly, as summative assessment results are not useful for informing daily instruction. They may, however, inform program and curriculum evaluation, which may in turn affect later instruction.

Critical Element 6.4 - Reporting

Notes
H. For example, interpretive guidance provided in conjunction with online data analysis websites, computer/tablet/smartphone data analysis applications, spreadsheet (e.g., Excel, Numbers, Google Docs) data analysis templates, user-friendly data file codebooks, and instructions for suggested analyses in common commercial or open-source data analysis software (e.g., Excel, SPSS, Stata, SAS, R).