February 2017

CCSSO
Council of Chief State School Officers

Evaluating Alignment in Large-Scale Standards-Based Assessment Systems

A white paper commissioned by the Technical Issues in Large Scale Assessment State Collaborative on Assessments and Student Standards of the Council of Chief State School Officers

Ellen Forte
edCount, LLC

THE COUNCIL OF CHIEF STATE SCHOOL OFFICERS

The Council of Chief State School Officers (CCSSO) is a nonpartisan, nationwide nonprofit organization of public officials who head departments of elementary and secondary education in the states, the District of Columbia, the Department of Defense Education Activity, and five U.S. extra-state jurisdictions. CCSSO provides leadership, advocacy, and technical assistance on major educational issues. The Council seeks member consensus on major educational issues and expresses their views to civic and professional organizations, federal agencies, Congress, and the public.

One Massachusetts Avenue, NW, Suite 700 • Washington, DC 20001-1431
Phone (202) 336-7000 • Fax (202) 408-8072 • www.ccsso.org

CONTENTS

LIST OF EXHIBITS

EVALUATING ALIGNMENT IN LARGE-SCALE STANDARDS-BASED ASSESSMENT SYSTEMS

Large-scale academic assessments have played a dominant role in U.S. federal and state education policies over the past couple of decades. As a result, the education measurement community and its methods have been catapulted into an arena where they are simultaneously hailed and denounced and where existing theory and technology do not always keep up with the demands of those requiring the tests. These circumstances have incentivized great progress in areas such as online and adaptive testing, the detection of cheating, and automated scoring for constructed-response items. They also obligate stock-taking of approaches to validity evaluation as the high-stakes requirements on score interpretation and use increases.

Among the many validity issues that presently concern test users is the evaluation of alignment among large-scale assessments and the academic content and performance standards on which they are based. In the pages that follow, we describe the current peer review expectations for alignment evidence, then present alignment as a problem of coherence to be addressed within policy and measurement contexts, describe popular approaches to evaluating alignment, and offer new perspectives that may yield more and more useful alignment information for test users and others.

ALIGNMENT EVIDENCE IN THE CURRENT PEER REVIEW GUIDANCE

Before diving into the current peer review guidance to lay out its expectations related to alignment, it's important to clarify that federal peer review is not the sole source of test users' obligation to gather and evaluate alignment evidence. This obligation really stems from the responsibilities defined in the *Standards for Educational and Psychological Testing* (hereafter referred to as "*The Standards*;" AERA/APA/NCME, 2014); the peer review guidance is based in large part on *The Standards*, but does not address *The Standards* comprehensively. Thus, while peer review may currently be the most salient call for alignment evidence, *The Standards* represent the true criteria for quality.

The current peer review guidance is the third version of such guidance; the U.S. Department of Education (ED) developed the first version in the late 1990s to structure the first ever federal peer reviews of states' systems of standards, assessments, and accountability as required under the *Improving America's Schools Act of 1994* (IASA, the 1994 reauthorization of the *Elementary and Secondary Education Act of 1965* or ESEA). Those rounds of peer review began in 1999 and continued even after enactment of the next reauthorization of ESEA, the *No Child Left Behind Act of 2001* (NCLB). ED developed new peer review guidance based on the NCLB legislation and began NCLB-related standards-and-assessment peer reviews in 2005.[1] That second version of the peer review guidance was updated a few times and continued to represent federal expectations until late 2015 when ED released the current version. Subsequently, President Obama signed the *Every Student Succeeds Act* (ESSA) into law in 2015 and states should expect some revisions to the peer review guidance so it reflects current legislation.

1 The NCLB peer review guidance separated accountability into a separate document and ED conducted accountability peer reviews in the spring of 2003.

Like its predecessors, the present peer review guidance is meant to communicate ED's expectations for how states are to demonstrate that their systems of standards and assessments meet the requirements of the ESEA legislation. Federal expectations related to alignment are outlined in a number of peer review elements; these appear in Exhibit 1.

Exhibit 1. 2015 Peer Review Elements Related to Alignment

2.1 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 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.

2.2 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.

3.1 The State has documented adequate overall validity evidence 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.

3.2 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.

3.3 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.

3.4 The State has documented adequate validity evidence that the State's assessment scores are related as expected with other variables.

4.3 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.

4.5 If the State administers multiple forms 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 and yield consistent score interpretations such that the forms are comparable within and across school years.

4.6 If the State administers assessments in multiple versions 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;

6.3 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.

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.

The list of elements in Exhibit 1 likely includes more elements than one might expect for a focus on alignment. However, alignment is about coherent connections across various aspects within and across a system and relates not simply to an assessment, but to the scores that assessment yields and their interpretations (Forte, 2013a, 2013b). The real alignment question is

> To what extent has the assessment and its operational system been designed to yield scores that reflect students' knowledge and skills in relation to the academic expectations defined in the standards, and how well has this design been implemented?

Before embarking on how to answer this question and address each of the alignment-related peer review elements, we need to ground the construct of alignment in both educational policy and educational measurement contexts so that we understand why it makes sense to view alignment broadly.

ALIGNMENT AS A CONSTRUCT AND GOAL IN POLICY AND MEASUREMENT

In this section, we consider the policy context for alignment evaluation, the measurement context, and the intersection of the two.

The Policy Context

The notion that large-scale academic assessments ought to be well-aligned with academic standards stems from the model of systemic reform (Smith & O'Day, 1991) on which U.S. federal education policy has been based since the 1994 reauthorization of ESEA (Forte, 2010). ESEA did not have a testing requirement initially, but subsequently included an evaluation requirement based on norm-referenced test scores that led to the invention of normal curve equivalent scores (Cross, 2003). These testing and evaluation expectations applied only to those schools getting Title I funding. Thus, the idea was that general assessments would yield good enough information to evaluate some form of program effectiveness and that comparability of outcomes between Title I and non-Title I schools was not important.

The publication of *A Nation at Risk* (1983) reignited concerns about equity and excellence in U.S. schools. The first discernable step toward addressing these concerns within federal education policy was the Hawkins-Stafford amendments to ESEA in 1988, which began an ongoing trend to loosen the limits on what qualifies a school to implement a school-wide Title I program. School-wide programs have much more flexibility in the use and combining of funds to serve all students whereas the alternative program type, Targeted-Assistance, meant Title I funds could only be used to support the lowest performing students, typically via a pull-out model.

In 1994, the Clinton administration introduced the Improving America's Schools Act of 1994 (IASA) as its version of ESEA. IASA mandated content standards, performance standards, assessments aligned with those standards, calculation of adequate yearly progress based on the aligned assessments, and again lowered the limits for schools' implementation of school-wide programs. In addition, IASA required states to apply the same sets of standards, assessments, and accountability systems to all schools rather than one approach for schools receiving Title I funds and another for other schools. Thus, systemic reform was now the official model for improvement in American schools and continues to be so today through NCLB and now ESSA.

The fundamental principle of the systemic reform model (Smith & O'Day, 1991) is coherence: one clear, common set of academic content expectations to drive both assessment design and curriculum and instruction; one set of clear, common expectations for performance in relation to those content standards; one set of assessments to measure performance in relation to those standards; and one model for evaluating progress. With these components in place, local educators could have the flexibility necessary to innovate and focus on achieving outcomes rather than on strictly adhering to the compliance model that targeted inputs.

Here is the policy target for alignment — The same expectations must drive (a) assessment and (b) curriculum and instruction. Assessments are meant to address the breadth and depth of the content and performance standards. Nowhere in any version of the ESEA legislation has the law required that every standard be reflected on the assessments.

The Measurement Context

The Standards for testing (AERA/APA/NCME, 2014) demand that test users establish a body of validity evidence to support the interpretation and use of each test score reported. The current edition of *The Standards* underscores this obligation in its very first standard:

"**Standard 1.0**. Clear articulation of each intended test score interpretation for a specified use should be set forth, and appropriate validity evidence in support of each intended interpretation should be provided." (AERA/APA/NCME, 2014, p. 23)

Like its predecessors, the 2014 *Standards* point to five sources from which test developers and evaluators ought to collect the evidence that is considered in a validity evaluation:

1. **Content** – evidence related to how well the assessment items and the assessment as a whole reflects the intended content domain

2. **Cognitive processes** – evidence related to how well the assessment items elicit the intended cognitive processes as students encounter, interpret, and respond to items and tasks on the assessment

3. **Internal structure** – evidence related to how well the scores an assessment yields related to one another in ways that correspond to expected inter-relationships among aspects of the intended content domain

4. **External relationships** – evidence related to how well the patterns of relationships between assessment scores and scores or other data elsewhere correspond to expected relationships between the assessment scores and outside criteria

5. **Consequences** – evidence related to how well decisions and actions based on the assessment scores or in anticipation of the assessment correspond to intended decisions and actions

These sources of evidence are neither types of validity nor discrete boxes on a checklist. They are where we look to find clues to help answer validity questions. Determining how well a test is aligned with the standards on which it is meant to be based would require consideration of various types of evidence from the sources above to address a number of validity questions. In addition, alignment can be understood as relating to score interpretation and use, not to an assessment in and of itself. Alignment is of utmost importance because without alignment evidence — from across multiple sources — it would be impossible to interpret assessment scores in relation to the standards on which those assessments are meant to be based. From this vantage point, we now step back to see how alignment has been framed in prior years.

Webb's Revolution

In his revolutionary article on alignment, which was funded by the National Science Foundation and involved a Task Force of 19 national experts on educational standards and measurement, Norman Webb defined alignment as

> "…the degree to which expectations and assessments are in agreement and serve in conjunction with one another to guide the system toward students learning what they are expected to know and do." (Webb, 1997, p. 4)

Further, Webb noted, "Alignment of expectations and assessments is a key underlying principle of systemic and standards-based reform" (Webb, 1997, p. 31). While he went on to suggest that "alignment corresponds most closely with content validity and consequential validity" (p. 4), subsequent developments in validity theory would now have us characterize alignment as depending heavily on evidence from content and consequential sources, though not to the exclusion of evidence from the other three sources.

Webb (1997) offers an informative synopsis of the state of states' standards at the time, observing that states' approaches to standards varied widely in terms of specificity (e.g., strand-level only, strands and performance indicators, strands, indicators, and instructional guidance), content areas (e.g., some combined mathematics and science into a single framework), and grade levels (e.g., grade ranges, only some grades). He provides this summary in part to make the point that attempts to evaluate alignment between a state's standards and assessments would first have to wrangle with grain size differences in the two documents. This was not a small problem then and continues to be of concern today.

Webb (1997) goes on to lay out a comprehensive set of criteria for evaluating alignment quality. (See Exhibit 2; the term "standards" replaces his term "expectations.")

Exhibit 2. The Complete Set of Webb's 1997 Alignment Criteria[23]

I. Content Focus

- **A. Categorical Concurrence**: correspondence between the topics in the standards and the topics by which assessment results are reported
- **B. Depth of Knowledge Consistency**: ratings of most cognitively demanding assessment activity for a topic within the standards as determined by number of ideas integrated, depth of reasoning required, knowledge transferred to new situations, multiple forms of representation employed, and mental effort sustained correspond to the same type of ratings of most cognitively demanding assessment activity for that same topic within the assessment
- **C. Range of Knowledge Correspondence**: standards and assessments cover a comparable span of knowledge within topics and categories[2]
- **D. Structure of Knowledge Comparability**: the relationships among ideas (e.g., no relationship, equivalent forms of the same idea, connection of many ideas within the content area, and connection of ideas within the content area and with applications to other areas) expressed in the standards are the same as those required to perform successfully on the assessments
- **E. Balance of Representation**: the weight by topic or subtopics in the standards corresponds with their weight on the assessments (weight could be determined by the proportion of activities by topic, proportion of average time allocated to do an assessment activity by topic, or according to some other rule)
- **F. Dispositional Consonance**: the desired dispositions toward the content area students are to develop as described in the standards are dispositional qualities are observed, monitored, and reported at designated levels within the system[3]

II. Articulation Across Grades and Ages

- **A. Cognitive soundness determined by best research and understanding**: the expressed or implied underlying theory of how students' learning progresses over time that is represented in the standards is reflected across grades in the assessments
- **B. Cumulative growth in content knowledge during students' schooling**: the expressed or implied understanding of how students' knowledge of content will be structured and will mature over time as represented in the standards is reflected across grades in the assessments

2 Webb (1997) firmly states that full coverage of the content within a set of standards "can be extremely difficult, if not impossible" (p. 18) and suggests states consider a form of matrix sampling that would sample content for each student and allow for a greater range of the standards to be covered at the school or district level. This approach was prohibited in the IASA and NCLB peer review guidance.

3 Webb (1997) does not suggest that evaluation of dispositional consonance is limited to large-scale assessments; rather, other aspects of the system should consider attitudes if they are included in the standards.

III. **Equity and Fairness**: students are afforded a fair and reasonable opportunity to demonstrate the full level of knowledge expected for all students. Assessment practices are such that variation of assessment results are only a variation in the attainment of expectations and free from being influenced by culture, ethnicity, gender, or any other irrelevant factor

IV. **Pedagogical Implications**

- A. **Engagement of students and effective classroom practices**: Instructional practices most likely to have students fully achieve expectations are the same as the instructional practices most likely to have students adequately demonstrate their attainment of these expectations on the assessments
- B. **Use of technology, materials, and tools**: adequate performance on assessments require students to be accomplished in using the full range of technology, materials, and tools as intended by the expectations

V. **System Applicability**: the public, teachers, students, and others within the system view expectations and assessments as closely linked, acceptable, attainable, and important

In light of what has become known as Webb's approach to alignment studies, the scope of the criteria in Exhibit 2 may seem surprisingly inclusive. Webb (1999) established a methodology for studying alignment that drew upon only 4 of his original 12 criteria (see Exhibit 3; Categorical Concurrence, Depth of Knowledge Consistency, Range of Knowledge Correspondence, and Balance of Representation).

Exhibit 3. The Four Alignment Criteria in the Webb Model for Evaluating Alignment (Webb, 1999)

1. **Categorical Concurrence**: the same or consistent categories of content appear in both the standards and the assessments
   - In practice: At least 50% of the categories (strands) within a set of standards are reflected among the assessment items, subscores correspond to the categories, and at least six items contribute to each subscore
2. **Depth of Knowledge Consistency**: what is elicited from the students on the assessment is as demanding cognitively as what students are expected to know and do as stated in the standards
   - In practice: As rated using a depth of knowledge (DOK) rubric, at least 50% of the items are at or above the DOK of the standard to which the items correspond
3. **Range of Knowledge Correspondence**: the span of knowledge expected of students by a standard is the same as or corresponds to the span of knowledge that students need to know in order to correctly answer the assessment items/activities
   - In practice: At least 50% of the standards within a category are represented by one or more items on the assessment
4. Balance of Representation: the extent to which items are evenly distributed across standards
   - In practice: The categories are relatively equally represented by items on the assessment

Webb and others (Beck, 2007; Bhola, Impara, & Buckendahl, 2003; Herman, Webb, & Zuniga, 2007; Webb, 2007) subsequently noted some weaknesses of the original methods and offered several recommendations for adjusting the Webb method that have been interpreted and implemented successfully in the District of Columbia, Florida, Georgia, and Puerto Rico.[4] First, the Categorical Concurrence criterion, which holds that the reporting categories for the assessment correspond with the categories into which the standards are organized and are associated with at least six items each, doesn't account for items that may appear on an assessment but aren't associated with any grade-level standards. This can be addressed through the use of a secondary Domain Concurrence criterion to represent the proportion of the items on the test that measure content defined in the grade-level standards. For example, if raters indicate that 67 percent of the items matched grade-level content at the standard or expectation level, the Domain Concurrence indicator would be 0.67. Of course, the goal would be 1 or 100 percent of the items linking to one or more standards.

Second, the 1999 Webb expectation that at least 50 percent of the items are at or above the depth of knowledge (DOK) of the associated standard could result in an assessment with a significant number of items that are more cognitively complex than the set of standards on which they are based. The alternative used in the states indicated above was a weighted index that requires consideration of the DOK ratings of the standards within a category and the distribution of DOK ratings for items associated with those standards. The first step in conducting a Webb-style alignment study involves panelists training to a DOK rubric and then assigning a consensus DOK rating to each standard for the grade and content area they are working on. Each panelist then works independently on an item-by-item review to (1) identify the standard that the item best reflects and (2) assign a DOK rating to the item.

Using the blueprint as the statement of how the standards are meant to be represented on the assessment, one can calculate an intended DOK at the category or strand level by averaging the standard-level DOK ratings within that category, weighted by the number of items associated with each standard in that domain. This intended DOK can then be compared with the DOK index that is calculated by averaging the DOK ratings for the items in that category.

Third, Webb's 1999 Balance of Representation criterion favors assessments that distribute score points equally across categories; if there are five categories in the standards, about 20 percent of the score points are expected to be associated with each of the five categories. Such standardized weighting may not make sense and could be quite imbalanced across categories for important curriculum-related reasons. Therefore, a suggested modification of this criterion involves the calculation of the proportions of score points associated with each category in the blueprint and then a comparison of those proportions with the proportions identified via the panelists' standard-rating of the items. This modification assumes that the blueprint has been developed to reflect the rational intention of the developers and this assumption must be checked.

4 These studies were commissioned by the state education agencies in these entities and conducted by edCount, LLC. The reports are available from the states or from edCount, LLC, with permission from the state/entity.

Other Alignment Study Models for General Assessments

Webb's 1999 alignment study approach was not the only method to emerge from his 1997 paper (Martone & Sireci, 2009). Achieve (2006) developed a model that addressed six criteria: accuracy of the test blueprint, content centrality, performance centrality, challenge, balance, and range. The accuracy of the test blueprint criterion requires that every test item corresponds to at least one standard. Content centrality addresses alignment between the content of the test item and the content of the related standard. Performance centrality criterion is a rating of cognitive complexity similar to Webb's DOK criterion. Challenge ratings encompass two factors — source of challenge (i.e., that the difficulty of the item is related to the content and not to construct-irrelevant factors) and level of challenge (whether the assessment includes a range of difficulty appropriate for the grade level). Balance relates to the degree to which the assessment and the content standards give similar emphasis to knowledge and skills within the content domain, and Range indicates whether the assessment items cover a representative sample of knowledge and skills from the content domain as defined by the content standards.

While the Achieve (2006) model and Webb's (1999) model clearly overlap in terms of what they attend to, they are very different in implementation. As described above, Webb's model involves panelists coming to consensus on the DOK ratings for each standard and then working independently to assign standards and DOK ratings to the items. Panelists look only at the standards and the assessments and do not discuss their item ratings. The two ratings for each item from each panelist are used in the analyses for the four criteria. The Achieve model is far more qualitative and relies heavily on the discussions among the expert panelists for insights into alignment quality. Achieve directly rejects the notion that alignment quality can be associated with numbers, "There is no mathematical formula for matching a test to standards" (Achieve, 2001, p. 11).

A second distinct difference between Achieve's approach and Webb's is that the Achieve model includes a review of the test blueprint in relation to the standards. That is, Achieve rightly situates the blueprint as the means by which the standards are translated into a framework for an assessment. The expected direct link between a set of standards and the items on a test in the Webb model ignores the role of the blueprint (or decision rules in adaptive designs) as the clear statement of what a test is meant to measure (Leighton & Gierl, 2007).

Like the Webb and Achieve models, the Surveys of Enacted Curriculum (SEC) alignment model (Blank, Porter, & Smithson, 2001; Porter, Smithson, Blank, & Zeidner, 2007) focuses on rating assessment items in relation to standards and cognitive demand. However, the SEC approach is notably different in several ways. First, the SEC model engages local educators in completing surveys of what is taught in the classroom so that what is taught (i.e., the enacted curriculum) can be compared in terms of topics and cognitive complexity with what is in the state standards. Second, SEC uses a common taxonomy of topics in each domain as the

basis for content ratings. If this taxonomy were in the leftmost column of a matrix, the state standards would be arrayed in a column to the right by topic. Items on the state test could be arrayed in another column as could educators' indications of what is being taught. In this way, the SEC model allows for comparisons across intended expectations for curriculum (i.e., the state standards), the assessed content, and the enacted curriculum. In addition, due to its use of a common taxonomy of topics, the SEC model readily supports comparisons across sets of standards. This negates the need to conduct the otherwise necessary two-way comparisons of one set of standards to conduct separate reviews to address both the 'How does set of standards A align with set of standards B?' and 'How does set of standards B align with set of standards A?' questions.

For peer review purposes, a state could choose to implement any of these models for alignment studies or any of a number of other variations on the Webb approach. Each has its strengths and the ultimate decision about which model to use should depend most on what information the state would find most helpful in improving the quality of its systems of standards and assessments in support of instruction.

Later in this document, we describe how the information gained from such studies would be used as evidence in a peer review package. In addition, we explain why the information from any of these studies is only a fraction of what represents adequate evidence of alignment.

Alignment Study Approaches for Alternate Assessments Based on Alternate Achievement Standards

Alternate assessments based on alternate achievement standards (AA-AAS) are those assessments designed for students with significant cognitive disabilities for whom the general assessment are inaccessible even with accommodations (U.S. Department of Education, 2005). AA-AAS must be based on the same content standards that apply to all students and which are used as the basis for the general assessments. The achievement standards[5] may be different from those associated with the general assessments so that they may more appropriately reflect the nature of how students with significant cognitive disabilities demonstrate their knowledge and skills.

The early days of federal peer review were also the early days of AA-AAS and the initial rounds of peer review revealed a striking lack of alignment evidence for these assessments even though the review criteria themselves were not particularly stringent (Forte, 2006) and did not even refer to AA-AAS. The NCLB peer review guidance, like the NCLB legislation, greatly ramped up the requirements for general and alternate assessments.

Tindal (2004) developed an alignment model specifically for AA-AAS that was based on the Webb (1999) approach. A key benefit of Tindal's approach was that it allowed for flexibility in

5 The term Academic Achievement Standards was introduced with the 2001 reauthorization of ESEA known as NCLB. The term Performance Standards was used to mean the same thing in the 1994 ESEA reauthorization known as IAS. Achievement standards and performance standards are the same thing as are achievement level descriptors (ALDs) and performance level descriptors (PLDs). States can use the terms they prefer.

how "items" were defined so it could more easily be applied to AA-AAS that were portfolio-based or otherwise observational in nature.

In developing the Links to Academic Learning (LAL; Flowers, Wakeman, Browder, & Karvonen, 2007) method for evaluating alignment among standards and alternate assessments for students with significant cognitive disabilities, researchers recognized the limitations of focusing on too few alignment criteria and wisely returned to the original Webb criteria (Webb, 1997). They also considered how others had built alignment evaluation methods that drew upon Webb (1997), and offered a methodology that includes ratings for eight criteria (see Exhibit 4).

The LAL method also took advantage of research on the nature of progressions in knowledge and skills across grades (Flowers et al., 2007). Reflected in LAL criterion 5, these progressions recognize important changes in content and skill expectations such that students master more deeply and apply more broadly the knowledge and skills in a domain over time and with instruction. Knowledge and skill expectations at lower grades may serve as prerequisite skills for those at higher grades and some new knowledge and skill expectations may be introduced at any grade (Webb, 2005). The LAL method does not recognize knowledge or skill expectations that are identical across grades as demonstrating necessary progression in expectations (Flowers et al., 2007). This criterion is particularly important for a model that evaluates alignment for the AA-AAS because it upends the common but misguided notion that students with significant cognitive disabilities cannot or do not learn.

The entirety of the LAL methods and criteria demand far more evidence of quality in and alignment among the standards, assessment, and instructional systems than any other alignment method. Application of the LAL method involves consideration of a large number of documents, such as information about the development of the AA-AAS, the blueprints, item specifications, the AA-AAS technical manual, examples of professional development for teachers about implementing the AA-AAS, policies about instruction and assessment for students with disabilities, and information about scoring the AA-AAS responses. Also unlike other methods, expert panelists engaged in a LAL study review students' responses rather than just the test form or a set of items and includes interviews with key personnel. Therefore, the LAL approach provides a far more comprehensive look at system alignment.

Further, LAL criteria 2 through 8 reject systems that blur distinctions across grade levels or otherwise lack a legitimate focus on grade-level academic knowledge and skills as defined in the sole set of content standards that are to apply to all students. An AA-AAS that is not based on grade-level academics would not fare well when evaluated via the LAL method.

Exhibit 4. The Eight Alignment Criteria in the Links to Academic Learning Alignment Method (Flowers, Wakeman, Browder, & Karvonen, 2007)

1. The Content is Academic
    a. Are the standards or extended standards on which the assessment is meant to be based academic in nature?
    b. Do the assessment items reflect the academic nature of the standards?
2. Referenced by Grade Level
    a. Are the standards or extended standards on which the assessment is meant to be based linked to a student's assigned grade level?
    b. Are the assessments items linked to academic content at the student's assigned grade level?
3. Fidelity with Grade Level Content and Performance
    a. To what extent to the assessment items reflect the content specified in the standards or extended standards?
    b. To what extent to the assessment items reflect the type and level of performance specified in the standards or extended standards?
4. The Content Differs in Range, Balance, and Depth of Knowledge (DOK)
    a. Categorical Concurrence: Are the categories of content in the assessment consistent with those in the standards?
    b. DOK Consistency: Is the cognitive complexity in the assessment items consistent with those in the standards?
    c. Range of Knowledge Correspondence: Is the span of knowledge necessary to answer the assessment items correctly consistent with the span of knowledge represented in the standards?
    d. Balance of Representation: To what extent are the assessment items evenly distributed across the categories in the standards?
5. Differentiation across Grade Levels or Grade Bands: Are the items age appropriate and across grades do the content and skill expectations change such that higher grades reflect new skills, broader applications, or deeper mastery of skills from earlier or that skills in earlier grades are prerequisites to those at higher grades?

6. Expected Achievement of Students is Grade Referenced Academic Content: Is there evidence that the student learned the content such that (a) there is evidence the student did not already have the skill, (b) the skill is performed without teacher prompting, and (c) the skill is performed across materials/lessons to show mastery of the concept versus rote memory of one specific response?

7. Barriers to Performance: Does the assessment allow for the widest range of students with significant cognitive disabilities possible to demonstrate what the know and can do such that (a) effective accommodations are allowed and (b) the assessment includes some items that do not require symbolic communication.

8. Instructional Program Promotes Learning in the General Curriculum: Is there evidence that the students who take the AA-AAS have meaningful access to the general curriculum including (a) professional development materials, (b) implementation of best practices as measured the Program Quality Indicators Checklist, and (c) grade-level standards-based instruction as measured by the Curriculum Indicators Survey.

Summary of Current Approaches to Alignment

The preceding descriptions of several approaches to alignment demonstrate that approaches share some similarities, but also reflect significant differences. In terms of the similarities, all methods consider both content and cognitive complexity and all involve reviews of standards and items by multiple individuals who have some demonstrated expertise in the content area or in curriculum and instruction for the target student population.

Exhibit 5. A Comparison of Several Current Approaches to Alignment

Method	Assessment Type	Standards and Assessments	Blueprints	Other policy and technical documents	Documents related to instruction
Webb (1999)	General	✓			
Achieve (2006)	General	✓	✓		
SEC (Blank et al., 2001)	General	✓			✓
Tindal (2004)	AA-AAS	✓			
LAL (Flowers et al., 2007)	AA-AAS	✓	✓	✓	✓

The Achieve and LAL approaches stand out for their consideration of blueprints and the LAL method stands alone in its inclusion of a host of other documents related to the complete system of standards, assessments, and instruction as implemented. For these reasons, the

Achieve and LAL methods are better aligned with Webb's original definition of alignment (Webb, 1997) than are other methods, including Webb's own approach (Webb, 1999). In addition, the Achieve, SEC, and LAL methods can yield particularly useful information about the standards and assessments as they are implemented in a coherent system that also includes instruction.

Any of these approaches could be used to address some of the alignment-related elements in the peer review guidance. However, in addition to the application of one of these study methods, a state would need to provide additional information related to alignment issues not addressed by these studies.

In the next section, we consider what we have learned about alignment evaluation, place the alignment study methods described above within a validity evaluation framework, and offer suggestions for how to gather additional evidence that may be helpful in both the peer review process and in a state's self-review of its assessments.

ALIGNMENT EVALUATION

Alignment as intended in U.S. federal education policy relates to coherence among the systems of standards and assessments that are meant to drive and allow evaluation of school improvement (Flowers et al., 2007; Forte, 2010; Webb, 1997). Alignment as operationalized in current practice tends to mean a mapping of individual content standards to test items on a test form and vice versa.

But, alignment evaluation is more than that and peer review demands more evidence than what alignment studies alone can provide. States may wish to consider additional information about their systems as they gather alignment evidence for peer review or any other purpose. In doing so, it is important to keep in mind that alignment relates to the connections between scores and domains represented by standards; it is not a characteristic of a test, per se. This parallels the primary tenet of validity — validity relates to score interpretations and uses rather than to assessments. Assessments are tools that we must design and implement carefully only because we need to count on the scores they provide. There is no reason to care about alignment unless it is in service of score interpretation and use.

When evaluating alignment, one should consider each of several key components in the standards-based assessment development process:

- Standards and the measurement targets they underlie
- The item and test development process
- Assessment blueprints
- Performance level descriptors

In terms of alignment logic, one begins with a set of standards[6] and determines what measurement targets assessment scores are meant to reflect. In the past couple of decades, many have turned to the principles of evidence-centered design (ECD; Mislevy & Haertel, 2006; Mislevy, Steinberg, & Almond, 2002) to guide this process. Under ECD principles, one induces a set of measurement targets that drive all subsequent test and item development. Where ECD or other principled-design approaches are not used, a test developer must find another way to clarify what is being measured by the test and by each item within it. A test developer must describe whatever item development model they use and justify why and how this model will produce items that address the claims.

Assessment blueprints, broadly defined, specify how tasks and test items are sampled from the item bank, how these are combined into the set that a student experiences as "the test," and how these tasks and items contribute to test scores. As such, blueprints should reflect the content and skill expectations of each task and item, as well as the full set of tasks and items "on the test" in relation to the complete set of claims. This role is relatively obvious for linear assessments; for adaptive assessments, developers must still identify the characteristics of the tasks and items that could be presented to a student during a test administration and how the combination of tasks and items is selected such that they could yield scores associated with the full set of claims. Thus, blueprints are concrete statements of what the test is meant to measure (Leighton & Gierl, 2007).

Blueprints or other types of specifications for what the test is measuring are necessary for both alignment and comparability. They indicate how the claims are represented on a given form or instance and support comparability across forms/instances, students, time, and sites.

Performance level descriptors (PLDs) are the statements that represent score meaning. Although standards-based achievement tests yield many types of scores (e.g., total test scale scores; sub-test raw or scale scores; various aggregations and disaggregations for student groups; aggregations at classroom, teacher, school, district, and state levels), the performance level is at the heart of all interpretations and uses in the standards-based context.

Taken together, we see how measurement targets, as initially defined by standards, should be intentionally and directly related to how items, tests, and the words that define test performance are developed. Now we consider how to evaluate the implementation and quality of the multi-step process by which standards are translated into assessments in ways that support intended score interpretations back to the standards.

6 Wise, Kingsbury, and Webb (2015) suggest that alignment to grade-level standards is not necessary at all for some tests, "…if the test is to be administered at multiple grades, or to be administered at different points in the school year, alignment of the test with the standards for the grade that the student is in may not be appropriate" (p.3). This is certainly not the case for the large-scale assessments required under ESEA, which must indeed be aligned with the standards for the grade in which the examinee is enrolled. Those who buy or commission other assessments, such as those that are administered at various points within school years and do not yield scores that are used for ESEA accountability purposes, may decide that alignment to grade-level standards is not important and ED has no standing to suggest otherwise. Even when a state's annual ESEA test is computer-adaptive in design, a state must demonstrate that scores at the student level meaningfully reflect the claims based on the standards, albeit with potentially fewer items.

Gathering Evidence

To be comprehensive and yield useful information for improving a state's system, an alignment evaluation would ideally address both the development process and the outcomes of the development process. Alignment studies like those described above address aspects of outcomes, but do not address process. To evaluate the development process requires some document collection and review.

Process is important to evaluating alignment because outcomes alone are limited in generalizable meaning. Consider the typical alignment study in which panelists review a set of items and connect them to standards and DOK rubrics. Even if that particular set of panelists found that particular set of items to reflect the standards rather well, that alone is insufficient evidence of whether other forms and items will be sound and aligned. Alignment evaluation must consider the system, not simply one example of a test. The question is whether the logic behind the development process is clear and reasonable such that it could yield aligned forms reliably.

To answer these questions, a state may wish to commission independent experts to collect and review development evidence and determine if the development process seems reasonable and likely to yield assessments that provide scores that can be interpreted as intended. The questions such a review would address are as follows:

1. What is the logic behind the construction of items and tests to support scores that can be interpreted and used as intended?
    a. How were measurement targets developed to reflect the standards?
    b. How were item development specifications developed to reflect the measurement targets?
    c. How were the blueprints developed to reflect the measurement targets?
    d. How were the PLDs developed to reflect the measurement targets?
2. Is this logic sound in terms of standards for professional practice and comprehensiveness without major gaps between steps in the logic chain?

The documents included in such a review would include at least the following:

- Standards documents and documents describing how the standards were developed
- Reports on the development of claims and measurement targets
- Reports on item and test development
- Reports on the development of the blueprints and other aspects of test design
- Reports on the development of the PLDs
- Test development staff resumes
- Documents used to train and guide item writers on content, bias and sensitivity, and any other aspect considered in item development

- Scoring guidelines
- Reports on the processes and outcomes of content and bias and sensitivity reviews
- Reports on cognitive labs, cognitive interviews, and any other means by which items were evaluated during development
- Pilot- and field-test reports

The Standards (AERA/APA/NCME, 2014) firmly call for this type of review and evaluation of these types of documents, which manifest the logic model that underlies the assessment endeavor. For example, *The Standards* demand "an analysis of the relationship between the content of a test and the construct it is intended to measure" (p. 15) in part via clear specifications of the construct(s) that the test is intended to assess (Standard 1.1). Test developers should "document the extent to which the content domain of a test represents the domain defined in the test specifications" (p. 89) and document the extent of this content representation by providing information about the design process.

Further, test developers and publishers should 1) "document steps taken during the design and development process to provide evidence of fairness, reliability, and validity for intended uses for individuals in the intended examinee population" (p. 85) and 2) "are responsible for developing and providing accommodations, when appropriate and feasible, to remove construct-irrelevant barriers that otherwise would interfere with examinees' ability to demonstrate their standing on the target constructs" (p. 67).

These types of evidence are necessary to help answer the fundamental validity questions about whether scores may be interpreted and used as intended. Has the system been built to yield interpretable scores? If not, the outcomes of external, post hoc alignment studies are not interpretable either. Perhaps the greatest benefit to process evaluation, however, is that the external evaluators can begin their work very early in the development timeline and provide formative feedback to test developers. It is far better to discover weaknesses in the foundation before the rest of the house is built.

Special Cases: Assessments Based on Merged Item Sets or Adopted Forms

As states consider the somewhat broader range of assessment options available to them under the 2015 ESEA reauthorization, some will choose to create tests by combining items from different sources or to adopt extant assessments that cannot be readily modified or augmented. Before launching into how a state addresses alignment in these cases, we note that the ESEA peer review requirements relate only to those assessments that yield scores that states use as the standards-based test for ESEA accountability purposes. However, *The Standards* apply to any assessments that yield scores used for any purpose; test users (e.g., state and local education agencies) are obligated to establish evidence in support of each interpretation and each use of each test score for each test they require so the same basic alignment evaluation principles apply.

Alignment Evaluation for Amalgamated Assessments Based on Multiple Item Sources

Amalgamated assessments based on multiple item sources are those that involve (a) the augmentation of an existing assessment through the addition of other items that contribute to students' scores or (b) the creation of entirely new assessments that draw from at least one item bank that was adopted in whole or in part by the test user post-development. In both cases, the focus for alignment evaluation is on the set of items that contribute to the scores that are intended to reflect the knowledge and skills defined in the standards. Therefore, the test user must start by establishing claims about what the scores from the compiled assessment are meant to mean just as they would do if they were building their own items from 'scratch.' This requires involvement of content and measurement experts who work together to define the major measurement targets based on the standards. Perhaps the same experts or some over-lapping group of them along with other stakeholders then engage in the development of PLDs; the PLDs and the claims and measurement targets drive development of blueprints.

The logic behind this process is the same as for tests built from the ground up, only starting in the middle rather than the beginning of the development process. One must still clearly state the purposes of the test scores, articulate how the test scores are meant to be interpreted and used (which is reflected in the PLDs), and build the tests to generate scores than can be interpreted in these ways. Sampling items from existing item banks can be a good way to save both time and money, but not if this approach gives short shrift to the internal aspects of validity. Thus, blueprints, which must be based on the PLDs and the claims and measurement targets, become the de facto representation of the measured domain. Items that populate forms must be selected and reviewed carefully by experts who were not part of the development process and are not employed by or represent any vendor involved in item development or who may profit from either the acceptance of existing items or the need to develop more items to fill gaps. No matter how extensive the combination of items from existing assessments or item banks may appear to be, one cannot assume they adequately address the claims and PLDs or fit the slots the blueprints comprise. The test user should evaluate each item considered eligible for placement on a test form, as defined by the blueprint, to ensure that it meets the criteria for what it is meant to measure.

To save both time and money in this process, the test user could choose to over-sample from the item banks to generate enough items to populate the forms for the first administration plus an additional 25 percent to 50 percent to allow for item substitutions based on content or sensitivity/bias reviews. Of particular importance to the amalgamation situation is that the item reviews require panelists to identify rather than confirm content matches. If done independently and involving an identification process, these ratings could be used in alignment study analyses rather than having to reconvene other panels to provide such ratings.

In addition to development of measurement targets, blueprints, and PLDs and the review of items, the user of amalgamated tests will need to collect evidence of the quality of the item development process; such evidence will be found in existing documents or from those

involved in the item development process. Not having been involved in the item development process is no excuse for not having evidence of item quality. Those selling item banks are obligated to provide comprehensive information about how items were developed to reflect what the sellers say they do.

Alignment Evaluation for Commercial Assessments

ESSA expressly allows for school districts to adopt existing, nationally-recognized tests with state approval, and nearly half of U.S. states already require all high school students to take such tests (Madda, 2016). However, if these commercially-available tests were not developed to reflect the common core state standards or states' content standards or to yield scores that are interpreted in relation to these standards, then use of these assessments may put states in the awkward position of having permission to use tests that were never meant to reflect achievement in relation to their state content standards.

There are at least two non-orthogonal approaches to the problem this poses for states that wish to use these tests and no other academic assessment in language arts or math at the high school level:

1. A state could try to create an amalgamated assessment based on the commercially-available test plus some other items.

2. A state could establish college- and career-ready achievement standards that reflect performance on the commercially-available test and have a loose association with some academic standards at the high school level.

The first approach could take a couple of forms. A state could try to determine how the commercially-available test items array against their claims or against a straw-man blueprint based on those claims and then develop or adopt items to fill the gaps. The state would have to determine how to administer the additional items in ways that do not disrupt the intact administration of the commercially-available test as this could have a major impact on the primary interpretation and use of those scores for college application purposes.

The logistical and cost challenges associated with trying to squeeze the commercially-available test pegs into holes they weren't meant for may make augmentation untenable for some states who plan to use these tests; such states would likely be better served by the second approach.

Approach two could also take a few forms. Perhaps the most useful and logistically plausible would be for the organizations that offer commercially-available tests to work with panels made up of stakeholders selected from across the country to develop PLDs based on the extensive empirical data these organizations have gathered and analyzed over the years on the associations between scores and college performance. Using a method applied in standards confirmation studies, panels could examine the actual items that correspond to empirically-derived score ranges such that students who scored in those ranges had a 67 percent chance of getting the items correct while students who scored in the next lower

range had less than a 33 percent chance of getting those items correct (Haertel & Lorie, 2004; Kane, 1994; Perie, 2008). Based on the items in these ranges, the panelists could develop PLDs that reflect the knowledge and skills that students who score in these ranges appear to demonstrate.

States that wish to use commercially-available tests for accountability purposes would be well-served to convene panels of local stakeholders, like those who would be involved in traditional standard-setting panels, to review these PLDs. These panels could be directed to provide feedback on the vendor-based PLDs and associated cut scores or allowed to make significant revisions to them in light of their state content standards. This process would not yield the type of alignment evidence generally expected of standards-based assessments, but could allow states to connect the commercially-available tests to claims about college- and career-readiness that draw more loosely from state content standards, per se.

ADDRESSING THE 2015 PEER REVIEW CRITERIA FOR ALIGNMENT EVIDENCE

The list of alignment-related peer review criteria presented in Exhibit 1 is now considered in relation to how a state would provide evidence to support the quality of its assessment system (see Exhibit 6). Although there are a number of alignment-related elements, there are actually very few underlying questions and many of these questions are repeated several times across peer review elements. A state should only respond to the question once; for subsequent elements, the state should refer to the answer and say why that response addresses the subsequent element.

Exhibit 6. Using the Comprehensive Alignment Evaluation Framework to Address the 2015 Federal Peer Review Criteria

Criterion Translation into *"Questions to Answer"* in Italics	**Basic evidence**	**Strengthened evidence**
2.1 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 of the State's academic content standards, and includes…(see bullets in the rows below): *What is the state's overall approach to building assessments that are aligned with standards?*	• Provide an illustration of the logic model that guided test design and test and item development • Describe this model in words	

Criterion Translation into *"Questions to Answer"* in Italics	**Basic evidence**	**Strengthened evidence**
• Statement(s) of the purposes of the assessments and the intended interpretations and uses of results *How are the assessment scores meant to be interpreted and used?*	• Refer to exact locations in evidence documents where the purposes of the assessments and the intended interpretations and uses of results are indicated	
• 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 *How were the blueprints (or other means by which operational items are selected for presentation to students) developed such that they reflect the standards?* *Is this process reasonably likely to produce assessments that could yield scores with the intended interpretations given professional standards for practice?* *Did this process produce a stable means by which assessments could yield scores with the intended interpretations?*	• Describe how claims and measurement targets were developed • Explain the blueprints, provide examples of them, and describe how the blueprints were developed to reflect the claims and measurement targets	• Independent evaluation indicating the process for developing claims and measurement targets was reasonable and sound • Independent evaluation indicating the process for developing blueprints was reasonable and sound • Independent evaluation of how well the claims and measurement targets reflect the standards • Independent evaluation of how well the blueprints correspond to the claims and measurement targets

Criterion Translation into *"Questions to Answer"* in Italics	**Basic evidence**	**Strengthened evidence**
• 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) *What was the state's approach to developing assessment tasks and items that reflect the full breadth and depth of the standards (via the claims and measurement targets)?* *Is this process reasonably likely to produce assessments that could yield scores with the intended interpretations given professional standards for practice?* *Did this process result in assessments that could yield scores with the intended standards-based interpretations?*	• Refer to previous response regarding the development of claims and measurement targets and blueprints • Describe how standards or claims and targets were translated into the means by which the items and assessments were developed • Independent evaluation of how well (a) the items and (b) the sets of items that contribute to students' scores reflect the claims and measurement targets in terms of categorical concurrence, DOK, range of knowledge, and balance of representation	• Refer to previous response describing independent evaluations of correspondence among standards, claims and targets, and blueprints • Independent evaluation of how items are developed to reflect the standards or claims and targets
• If the State administers computer-adaptive assessments, the item pool and item selection procedures adequately support the test design *How were item selection algorithms developed such that they could yield sets of operational items that reflect the standards?* *Is this process reasonably likely to produce assessments that could yield scores with the intended interpretations given professional standards for practice?* *Did this process produce a stable means by which assessments could yield scores with the intended interpretations?*	• Describe how the adaptive algorithms are designed so that the complete set of items presented to a student and on which students' scores are based reflect the breadth and depth of the claims and measurement targets • Independent evaluation indicating the process for item selection is reasonable and sound	• Independent evaluation of how well (a) the items and (b) the sets of items that contribute to students' scores reflect the claims and measurement targets

Criterion Translation into *"Questions to Answer"* in Italics	**Basic evidence**	**Strengthened evidence**
2.2 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 *This element repeats previous concepts about item and test development. Because this element refers to achievement, we take that as the opportunity to provide evidence that the assessments have been developed to reflect the full range of performance as defined in the PLDs. Thus, the only new questions here are* *What was the state's approach to developing PLDs that reflect the full breadth and depth of the standards (via the claims and measurement targets)?* *Is this process reasonably likely to produce assessments that could yield scores with the intended interpretations given professional standards for practice?* *Did this process result in assessments that could yield scores with the intended standards-based interpretations?*	• Refer to previous response regarding the development of claims and measurement targets, blueprints, and how claims and targets were translated into task models, item templates, and other means by which the items and assessments were developed • Describe how claims and targets were translated into PLDs • Describe independent evaluation indicating the process for developing the PLDs was reasonable and sound	• Refer to previous responses describing independent evaluations of each step in the chain from standards to claims and targets to items and to tests • Independent evaluation of how well the PLDs reflect the claims and measurement targets
3.1 The State has documented adequate overall validity evidence 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:		

Criterion Translation into *"Questions to Answer"* in Italics	**Basic evidence**	**Strengthened evidence**
• 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; *This element repeats previous concepts about alignment to standards in terms of breadth and depth, highlighting the evidence that the operational assessments are aligned* *Did this process result in assessments that could yield scores with the intended standards-based interpretations?*	• Refer briefly to previous responses regarding translations from standards to claims and measurement targets to task models and item templates and blueprints and PLDs • Refer to previous responses describing independent evaluations regarding categorical concurrence, DOK, range of knowledge, and balance of representation	
• 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 *This element repeats previous concepts about alignment to standards in terms of breadth and depth, highlighting the evidence that the operational assessments are aligned* *Did this process result in assessments that could yield scores with the intended standards-based interpretations?*	• Refer to previous responses regarding translations from standards to claims and measurement targets to task models and item templates and blueprints and PLDs • Refer to previous responses describing independent evaluations regarding categorical concurrence, DOK, range of knowledge, and balance of representation	

Criterion Translation into *"Questions to Answer"* in Italics	**Basic evidence**	**Strengthened evidence**
3.2 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	• Refer to previous response regarding how claims and targets were translated into task models and item templates that could yield items that demand the intended cognitive processes • Independent evaluation indicating the process for translating the claims and targets into task models and item templates that address the intended cognitive processes was reasonable and sound	• Describe evidence from any studies a state or vendor may have conducted to address cognitive processes. Such students could involve cognitive labs or cognitive interview methods, although these methods are expensive and may not provide reliable evidence about students' actual problem solving while taking assessments (Leighton, 2015; Padilla, Benitez, Herrera, & Rico, 2015)
3.3 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 *What was the state's approach to developing assessments that reflect the structure of the standards (via the claims and measurement targets)?* *Is this process reasonably likely to produce assessments that could yield scores with the intended interpretations given professional standards for practice?* *Did this process result in assessments that could yield scores with the intended standards-based interpretations?*	• Refer to previous descriptions of how claims and targets were developed and translated into task models and item templates (including scoring models) and blueprints	• Refer to previous responses describing independent evaluations of each step in the chain from standards to claims and targets to items and to tests • Refer to previous responses regarding evaluations of how well PLDs reflect standards and claims and targets • Describe evidence from any studies a state or vendor may have conducted to evaluate internal structure, perhaps using factor analysis methods

Criterion Translation into *"Questions to Answer"* in Italics	**Basic evidence**	**Strengthened evidence**
3.4 The State has documented adequate validity evidence that the State's assessment scores are related as expected with other variables *Given the nature of standards-based assessments, this is not a heavy burden. A state can provide adequate evidence to support valid interpretations and uses of its assessment scores without evidence related to external relationships. If a state does wish to collect this evidence, the question would be:* *Do the patterns of relationships with external indicators reflect the expected patterns?*	• There is no expectation that a state should consider relationships with external indicators while designing or developing a standards-based academic achievement assessment; thus, no process-related evidence is necessary	• Evidence from any studies a state or vendor may have conducted to address how state assessment scores may relate to other indicators of student achievement with strong caveats about the interpretability of the results. Other indicators at the student level could include grades or other achievement test scores; at the state level, NAEP scores could be considered other indicators. The state would need to explain how and why it would expect its assessment scores to relate to external indicators
4.3 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 *This element repeats prior concepts about the range of performance that students could demonstrate via the assessment. It is not asking about performance across different student groups.* *What was the state's approach to developing PLDs that reflect the full breadth and depth of the standards (via the claims and measurement targets)?* *Is this process reasonably likely to produce assessments that could yield scores with the intended interpretations given professional standards for practice?* *Did this process result in assessments that could yield scores with the intended standards-based interpretations?*	• Psychometric evidence relating to precision/ reliability • Refer to previous responses regarding the development of the PLDs	• Refer to previous responses regarding independent evaluations of PLD alignment

Criterion Translation into *"Questions to Answer"* in Italics	**Basic evidence**	**Strengthened evidence**
4.5 If the State administers multiple forms 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 and yield consistent score interpretations such that the forms are comparable within and across school years *This is a comparability question and relates to alignment in that the item development process and blueprint must support the ongoing alignment that is necessary for comparability. The alignment questions are as follows:* *What was the state's approach to developing assessments that reflect the full breadth and depth of the standards (via the claims and measurement targets)?* *Is this process reasonably likely to produce assessments that could yield scores with the intended interpretations given professional standards for practice?* *Did this process result in assessments that could yield scores with the intended standards-based interpretations?*	• Psychometric evidence relating to comparability • Refer to previous descriptions of how claims and targets were developed and translated into task models and item templates (including scoring models) and blueprints	• Refer to previous responses describing independent evaluations of each step in the chain from standards to claims and targets to items and to tests
4.6 If the State administers assessments in multiple versions within a content area, grade level, or school year, the State:		

Criterion Translation into *"Questions to Answer"* in Italics	**Basic evidence**	**Strengthened evidence**
• Followed a design and development process to support comparable interpretations of results for students tested across the versions of the assessments *This is a comparability question and relates to alignment in that the item development process and blueprint must support the ongoing alignment that is necessary for comparability. The alignment questions are* *What was the state's approach to developing assessments that reflect the full breadth and depth of the standards (via the claims and measurement targets)?* *Is this process reasonably likely to produce assessments that could yield scores with the intended interpretations given professional standards for practice?* *Did this process result in assessments that could yield scores with the intended standards-based interpretations?*	• Psychometric evidence relating to comparability • Refer to previous descriptions of how claims and targets were developed and translated into task models and item templates (including scoring models) and blueprints	• Refer to previous responses describing independent evaluations of each step in the chain from standards to claims and targets to items and to tests • Refer to previous responses regarding PLD alignment
6.3 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 *This element is related to the underlying assumption that the standards do support development of the academic knowledge and skills necessary to be successful in post-secondary settings. The alignment related questions are* *What was the state's approach to developing assessments that reflect the full breadth and depth of the standards at each grade level?* *Is this process reasonably likely to produce assessments that could yield scores with the intended interpretations given professional standards for practice?* *Did this process result in assessments that could yield scores with the intended standards-based interpretations?*	• The primary response to this element must refer to how the state defines and evaluates success in post-secondary settings • Refer to previous descriptions of how claims and targets were developed and translated into task models and item templates (including scoring models) and blueprints	• This response requires evidence that would come from an evaluation of the standards and their connections with external definitions of or requirements for success in post-secondary settings. This evidence could come from studies of test performance in relation to performance in those settings. Evidence that the standards are related to indicators of post-secondary demands or outcomes may also be helpful • Refer to previous responses describing independent evaluations of each step in the chain from standards to claims and targets to items and to tests

Criterion Translation into *"Questions to Answer"* in Italics	**Basic evidence**	**Strengthened evidence**
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	• Refer to previous responses regarding translations from standards to claims and measurement targets to task models and item templates and blueprints and PLDs	• Refer to previous responses describing independent evaluations of how the claims and targets and PLDs for the AA-AAS reflect the standards

CONCLUSIONS

Alignment is about coherent connections across various aspects within and across a system (Forte, 2013a, 2013b). A comprehensive alignment evaluation must consider many elements, including the claims about what a test measures and how those claims are ultimately translated into test items and the tests that yield scores. All parts of the testing process must fit within the over-arching vision of what the assessment scores are meant to indicate about a student's knowledge and skills.

To support valid, standards-based interpretations of assessment scores, an alignment evaluation should encompass two parts: (1) a statement of how the system is adequately designed to support alignment and (2) how the system is built and implemented in ways that actually are aligned to support interpretations of students' scores in relation to the standards. In many ways, this approach is itself tightly aligned with current approaches to validity evaluation that require articulation of a logic model and the testing of assumptions and links within that model (Kane, 2006). Evidence that test content reflects the concepts that were meant to be measured is also one of the critical sources of information necessary to support valid interpretations of test scores (AERA/APA/NCME, 2014).

REFERENCES

Achieve, Inc. (2006). *An alignment analysis of Washington State's college readiness mathematic standards with various local placement tests.* Cambridge, MA: Author.

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

Beck, M. D. (2007). Reviews and other views: Alignment as a psychometric issue. *Applied Measurement in Education, 20*(1), 127-135.

Bhola, D. S., Impara, J. C, & Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice, 22*(3), 21-29.

Blank, R. K., Porter, A. C., & Smithson, J. L. (2001). *New tools for analyzing teaching, curriculum and standards in mathematics and science (Report from the Survey of Enacted Curriculum Project; National Science Foundation REC98-03080).* Washington, DC: Council of Chief State School Officers.

Cross, C. (2003). *Political education: National policy comes of age*. New York City, NY: Teachers College Press.

Flowers, C., Wakeman, S., Browder, D. & Karvonen, M. (2007). *Links for academic learning: An alignment protocol for alternate assessments based on alternate achievement standards*. Charlotte, North Carolina: University of North Carolina at Charlotte.

Forte, E. (2013a*). Re-conceptualizing alignment in the evidence-centered design context*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Forte, E. (2013b). *Evaluating alignment for assessments developed using evidence-centered design.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Forte, E. (2010). Examining the assumptions underlying the NCLB federal accountability policy on school improvement. *Educational Psychology, 45*(2), 76-88.

Forte, E. (June, 2006). *Lessons learned from peer review.* Paper presented at the Annual Large Scale Assessment Conference of the Council of Chief State School Officers, San Francisco, CA.

Haertel, E. H., & Lorie, W. A. (2004). Validating standards-based test score interpretations. *Measurement: Interdisciplinary Research and Perspectives, 2*(2), 61–103.

Herman, J. L., Webb, N. M., & Zuniga, S. A. (2007). Measurement issues in the alignment of standards and assessments: A case study. *Applied Measurement in Education, 20*(1), 101-126.

Kane, M. T. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*(3), 425–461.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.

Leighton, J. P. & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice, 26*, 3-16.

Madda, M. J. (February 8, 2016). *Why the SAT and ACT may replace PARCC and Smarter-Balanced.* Retrieved February 8, 2017, from https://www.edsurge.com/news/2016-02-08-why-the-sat-and-act-may-replace-parcc-and-smarter-balanced.

Martone, S, A. & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research, 79*(4), pp. 1332-1361.

Mislevy, R. J. & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice, 25*(4), pp. 6–20.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3-67.

Padilla, J. L., Benitez, I., Herrera, A., & Rico, J. (2015). *Complementarity between cognitive interviewing findings and DIF results: Enhancing validation and test design.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Perie, M. (2008). A guide to understanding and developing performance-level descriptors. *Educational Measurement: Issues and Practice, 27*(4), 15-29.

Porter, A. C., Smithson, J., Blank, R., & Zeidner, T. (2007). Alignment as a teacher variable. *Applied Measurement in Education, 20*(1), 27-51.

Sireci, S. G. & Martone, A. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research*, 79(4), 1332-1361.

Smith, M. S., & O'Day, J. (1991). Systemic school reform. In S. H. Fuhrman, and B. Malen (Eds.), *The Politics of Curriculum and Testing, 1990 Yearbook of the Politics of Education Association*. London and Washington, DC: Falmer Press, 233-267.

Tindal, G. (2004). *Alignment of alternate assessments using the Webb system*. Washington, DC: Council of Chief State School Officers.

U.S. Department of Education. (2005). *Alternate achievement standards for students with the most significant cognitive disabilities.* Washington, DC: Author.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (Council of Chief State School Officers and National Institute for Science Education Research Monograph No. 6). Madison: University of Wisconsin, Wisconsin Center for Education Research.

Webb, N. L. (1999). *Alignment study in language arts, mathematics, science, and social studies of state standards and assessments for four states*. Washington, DC: Council of Chief State School Officers.

Webb, N. L. (2005, November). *Alignment, depth of knowledge, and change*. Paper presented at the annual meeting of the Florida Educational Research Association, Miami, FL.

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education*, *20*(1), 7-25.

Wise, S. L., Kingsbury, G. G., & Webb, N. L. (2015). Evaluating content alignment in computerized adaptive testing. *Educational Measurement: Issues and Practices, 34*(4), pp. 41-48.

Council of Chief State School Officers

One Massachusetts Avenue, NW, Suite 700
Washington, DC 20001-1431
voice: 202.336.7000 | fax: 202.408.8072