Session 1B: Alignment

Panelists: *Meagan Karvonen (karvonen@ku.edu), Phoebe Winter (phoebe.winter@outlook.com), Brooke Nash (bnash@ku.edu), Zach Warner (zachary.warner@nysed.gov)*

[Session 1B Alignment slides 4–20]
After providing an overview of the session, Dr. Karvonen identified the primary CEs for alignment in peer review as follows: 2.1 (Test Design and Development); 3.1 (Overall Validity); 2.2 (Item Development); and 4.7 (Technical Analysis, Ongoing Maintenance). She noted that states often forget to address CE 4.7 in their peer review submissions. A separate session will focus on CE 6.3 (Challenging and Aligned Academic Achievement Standards). She mentioned that alignment may also affect other CEs (e.g., 3.3, 3.4, 4.3, and 4.5).

A state's test design and test development process should be well suited for the content, technically sound, and align the assessments to the depth and breadth of its academic content standards for the grade that is being assessed. This refers to the depth and breadth of the state's grade-level academic content standards in terms of balance of content (i.e., knowledge, cognitive process, and cognitive complexity). Dr. Karvonen stressed that for CE 2.1, everything flows from the statement of purposes and intended uses. Blueprints support test development, showing depth and breadth. Assessment tailored to knowledge and skills in the standards, include appropriately complex applications, and the CAT item pool and item selection procedures support test design.

Assessments measure the knowledge and skills in the content standards, including documentation of adequate alignment between the assessments and the content standards in terms of content (i.e., knowledge and process), balance of content, and cognitive complexity. States must also submit documentation that the assessments address the depth and breadth of the content standards. For item development, the state must show that it uses reasonable and technically sound procedures to develop and select items that assess student achievement based on the state's academic content standards in terms of content and cognitive process, including higher-order thinking skills. For technical analysis and maintenance, the state must demonstrate that it has a system for monitoring, maintaining, and improving (as needed) the quality of its assessment system. Evidence should include clear and technically sound criteria for the analyses of all of the assessments in its assessment system.

States often hold misconceptions about alignment. Staff members working on assessment may perceive that alignment is only about the relationship of items to content standards. But ED and subject matter experts consider it more broadly. Another misconception is that alignment evidence comes from an external study on an operational item or task pool. Although this is important, it is not the sole evidence. Many in the field think that peers only accept alignment evidence that uses Norman Webb's methods and criteria (1997), but there are many options and updates in methodology. Dr. Karvonen reviewed an example of how alignment would be poor in the peer review context if a state assumed Webb's criteria.

Dr. Karvonen provided a brief overview of the assessment methodology options discussed during the plenary session: (1) TYA; (2) portfolio, project-based, or performance assessment (noting that this is only part of a student's achievement assessment); and (3) matrix sampling of assessed content. She referred participants to the Multiple Approaches Handout, which was accessible via a QR code (see Framing Table for Multiple Approaches to Assessment Design). Multiple approaches to assessment may bring different ways of defining and evaluating alignment. They have different components to align, and their thresholds of "adequate" alignment must be based on test design and blueprint and intention. States will need a different view of alignment as it relates to validity. She stressed that states must be intentional from the beginning when selecting their approaches.

[Session 1B Alignment slides 21–34]
Dr. Nash noted that in the traditional view, content validity is aligned with an external study (typically using Webb's methodology). In contrast, the comprehensive view draws upon five sources of evidence for validity: (1) content; (2) cognitive process; (3) internal structure; (4) relationships with other variables; and (5) consequences. In this approach, claims about alignment within a coherent system require evidence within multiple sources of validity evidence. The comprehensive approach offers states the following benefits. It supports the development and implementation of well-articulated validity arguments and validation plans. This benefits states by making accumulating and synthesizing validity evidence for peer review much easier. Accumulating alignment evidence throughout the design and development process can provide opportunities for identifying potential alignment issues earlier on in the development process (i.e., in time to correct them prior to operational administration). Additionally, establishing alignment expectations as part of the test design supports external partners in designing and conducting alignment studies that are consistent with design. This also makes reconciling and addressing alignment findings easier. Finally, the comprehensive approach to alignment more directly supports score interpretations and uses than the traditional view. Dr. Nash suggested that states interested in a new approach to assessment consult Ellen Forte's 2017 white paper for background and examples. She reviewed potential alignment relationships, noting that states should consider evaluation alignment from a broad perspective early in the process.

For TYA/IE models, states should consider various potential design features when developing alignment plans. These include plans for scoring and reporting after each administration (embedded or interim) and at the end of the year (summative), blueprint specification within and across assessment windows, the size and scope of the item bank, and the year-round administration of short tests or testlets. Overall, states should consider the totality of assessments and what aspects meet the blueprint's requirements for summative purposes. States also should consider the types of evidence that could be used to demonstrate alignment of assessments with academic content standards for TYA/IE models and organize them by procedural and evaluative evidence types. States should also discuss procedural issues when describing their content structure and coverage of content, as specified by test blueprints. This section would include any flexibility in content selection and any effects on inferences that can be made from results.

States should also discuss how they developed achievement level descriptors (ALDs) and proficiency level descriptors (PLDs)—including clear articulation of student performance expectations achieved by the end of the year. They should also describe the score report (or reporting dashboard) design, scoring models, and procedures in a way that is consistent with their intended metrics for reporting. An important component of this section is formative alignment checks during the process. Evidence-centered design can be used to develop high-quality items aligned to targeted content, and states should train test item writers in this method. Peer review submissions should also include a description of the state's procedures for monitoring blueprint coverage. Dr. Nash stressed that states often collect evidence that they do not submit in peer review. It benefits states to include this information so the submission is as comprehensive as possible. States should include the results from educator review of items prior to field-testing and for the items selected for operational forms. They should also submit an analysis of blueprint coverage and an external alignment study. She reviewed DLM content structures, noting that external evaluators must understand these structures for alignment findings to be meaningful and achieve the goals of the assessment system.

[Session 1B Alignment slides 35–52]
Dr. Warner focused on performance assessment, which is a broad category. It must be valid but can be over-standardized in a way that stifles innovation and opportunities for students to show what they know and can do. He commented that the field has become efficient at large-scale assessment practice and has moved away from theory. He suggested taking a step back to consider what is being aligned and the overall purpose of assessment (e.g., goals and policy outcomes). The Webb method is only one approach to alignment. Performance assessments cover depth but not necessarily breadth of content. They can weave complexity throughout the varied tasks so that students have different entry points to the work. Research is needed on the complexity drivers in performance assessment. The administration time of performance assessments varies greatly, and scoring criteria show what is valued.

For performance assessments, procedural considerations include domain analysis or the "unpacking" of learning standards to identify the knowledge and skills that are best assessed by a performance task (as opposed to a selected response or another item). Careful planning and documentation are essential, and states must articulate the PLDs. States should include task specifications and blueprints (including any flexibilities and their effects on inferences) and information related to task development (e.g., training for writers and proctors). Regarding the evaluation of performance assessments, states should include the results from an educator review of tasks, an external alignment study, and a demonstration that flexibility within administration does not hinder the measurement of the intended content and processes. Dr. Warner remarked that in New York, teachers are fully engaged in performance assessment, which helps with the evaluative component. He added that failure of alignment often has to do with various interpretations of the learning standards and stressed the importance of agreement from the beginning of the process.

Matrix sampling addresses how the breadth and depth of the state's learning standards will be covered (1) within each form/year and (2) across forms/years to ensure full coverage. Under ESSA, states are required to cover all students during assessment and report grade-level proficiency across the standards. These requirements must be considered early in the process to ensure that the coverage supports the necessary level of reporting. Dr. Warner described how states can address potential peer reviewer concerns about matrix sampling. To explain why matrix sampling is appropriate for the assessment program, states should present clear arguments and describe the opportunities. They may include a performance-based standard that is tied to a theory of action. States should provide evidence that the blueprint can support everything that comes downstream at the district and school levels. States should show that students across the continuum of proficiency will receive an aligned test.

Procedural evidence for matrix sampling includes the rationale for the blueprint (i.e., the theory of action and claims). States should provide evidence that combined blueprints cover the breadth and depth of the learning standards. Peer reviewers need to see that states have implemented their blueprints and that students take an assessment with sufficient breadth of coverage each year. States should submit a description of the test development steps that promote alignment (e.g., task templates and item writer training). In this section, states should tell peers what they did and show how each step builds the case for alignment. Dr. Warner suggested that states educate peers about their assessment programs—their approaches to alignment, what they value, and how they demonstrate these things during assessments. The "chain of evidence" that states need to show to peer reviewers includes evidence of the following: (1) intended content relationships, (2) procedures, (3) an external alignment study with an appropriate design and criteria, and (4) how the state interprets and responds to findings.

Dr. Warner described potential components for each area of evidence, noting specific resources with guidance on various methodological options. He recommended that states generally explain the relevant content relationships each time they present the evidence—especially when an atypical design is used. States should support qualitative statements with quantitative data, which helps frame the data and make the case for alignment. States should also synthesize the evidence (e.g., procedural and empirical evidence from all stages) to make the case for how alignment goals are met in relation to peer review requirements. If relevant, they should make additional validity claims.

Participants applied information from the session in an activity on alignment and reported highlights of their small-group discussions.

Kevin O'Hair, Academic Program Manager at the Kentucky Department of Education, reported that his group discussed a recent internal review that involved an examination of breadth and depth, alignment, the range of cognitive complexity, and other aspects of the assessment system. Regarding challenges, Audra Ahumada, Deputy Associate Superintendent of Assessment at the Arizona Department of Education, mentioned the complexity of assessment of English learning proficiency (ELP). Her group concluded that a focus on PLDs provides strong evidence and a basis for developing the procedures for ELP assessment.

Questions and Comments

A staff member of the Wyoming Department of Education remarked on the ongoing changes to her state's assessment system. The team is challenged by defining depth and breadth in the context of a reduction in the number of standards. Dr. Warner replied that a single standard is problematic, but with more than one, states might articulate specific items and PLDs. New York has done this. He cautioned that the learning standards cannot be questioned under ESSA and suggested that the state ensure that its blueprints and documentation link to the standards. It will be crucial to provide teachers with guidance on the standards. The state might discuss depth and breadth in terms of consistency, clarify its definition, and link it to the blueprint—mentioning what will be done and how that connects with content. Dr. Winter suggested taking an approach that evaluates how the test matches each learning standard.