Session 1D: Overall Validity

Panelists: *Phoebe Winter (phoebe.winter@outlook.com), Chris Rozunick (christine.rozunick@tea.texas.gov), Zach Warner (zachary.warner@nysed.gov), Nathan Dadey (ndadey@nciea.org)*

[Session 1D Overall Validity slides 5–11]
After providing an overview of the session, Dr. Winter discussed considerations for overall validity for the multiple approaches to assessment. As validity is the most fundamental consideration for developing assessments, the information covered has implications for multiple CEs. These include 3.1 (Overall Validity), 3.2 (Validity Based on Cognitive Processes), 3.3 (Validity Based on Internal Structure), and 3.4 (Validity Based on Relations to Other Variables). Regarding overall validity (which includes validity based on content), the state must demonstrate that its academic assessments measure the knowledge and skills specified in the state's academic content standards. Dr. Winter suggested that states should focus on presenting evidence for this important global characteristic that is often overlooked. The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score and assessment results and interpretations. Validity is a process, and states should plan on monitoring it over the longer term and changing assessment based on what is learned.

The multiple approaches considered typically have purposes and uses that are different from or additional to traditional state summative assessments. To meet these purposes and uses, they may have between-student variation in the "what, when, and how" of student assessment. For example, in performance assessment examining students' cognitive processes, the "what, when, and how" could affect validity. She referred participants to the Multiple Approaches Handout. Dr. Winter emphasized that although the field often considers validity to be a property of the test, this is shorthand and not precisely correct. Rather, validity is a property of the proposed interpretations of test scores for specific uses and is a matter of degree.

[Session 1D Overall Validity slides 12–19]
Dr. Nash remarked that the concept of an argument-based approach to assessing the validity of assessments was developed in the 1980s and offers a framework for this process. The two steps to an argument-based approach to validity are to (1) state the claims associated with the proposed interpretation or use and (2) evaluate the claims. She noted that there are good resources for this process. For the first step, states need a logic model for making explicit the inferences, claims, and assumptions necessary to make links between the observed test score and intended interpretations and uses. The model should be clear, coherent, plausible, and comprehensive. Well-articulated interpretation and use arguments include the intended interpretations of results and the uses of results. Dr. Nash reviewed some examples.

When states use multiple approaches and proposed interpretations, they should consider that test scores can be interpreted in multiple ways and have multiple possible uses. It is also important to understand that the validity of a proposed interpretation or use depends on how well the evidence supports the proposed interpretation or use. Finally, more ambitious interpretations and uses require more evidence. Validity arguments provide an evaluation of all

the claims and assumptions outlined in the argument for the interpretation or use. Well-articulated validity arguments put the evidence within each of the claims and assumptions in the argument for the interpretation or use. They also evaluate the degree to which the evidence supports each claim and assumption and integrate the evidence and theory into a coherent argument. Dr. Nash reviewed some examples. Although a theory of action is not necessary, it can help states extend beyond these claims to include intended change in what is assessed and support a robust articulation of the use of assessments.

[Session 1D Overall Validity slides 20–25]
Dr. Warner commented that regardless of the framework used to explain and evaluate the validity argument, states can summarize the evidence of validity according to the peer review CEs. Establishing a strong validity argument supports all aspects of an assessment—driving design decisions, supporting item and task selections, connecting content with results, and guiding reporting decisions. He added that it can be overwhelming to think about validity deeply, but essentially it is making a claim and supporting it with evidence. The validity argument is essential and requires work upfront. Validity arguments help states avoid arbitrary decisions about assessment systems.

Dr. Warner used New York's use of performance-based items in science examinations at all levels as an example. This practice is popular and helps drive instruction. The underlying theory of action is as follows: If students are presented with opportunities to demonstrate course-specific or grade-level science knowledge and skills via hands-on activities, their performance will produce evidence of their comprehension of the specific learning standards associated with that knowledge and those skills and contribute to a total score that enables inferences about student attainment of the learning standards for the course or grade level. He noted that these characteristics are needed to meet the requirements of ESSA and guide instruction.

Theories of action for performance-based assessments require various types of validity evidence, including that the tasks are appropriate for the content. States can demonstrate this by providing peer reviewers with blueprints and content coverage and task and form specifications. Dr. Warner noted the importance of including complexity information for innovative tests or tasks. States should support the claim that the tasks are designed to solicit intended evidence (e.g., task development specifications and processes and alignment studies). States should explicitly show that the evidence produced by tasks informs the intended interpretations for all student groups.

[Session 1D Overall Validity slides 26–28]
Ms. Rozunick discussed overall validity in the context of the TTAP. Although this is a pilot program and may not go to peer review in the immediate term, considering this process will help the team build a validity argument. She reviewed TTAP's theory of action and listed the research areas that will support it. Through TTAP, Texas is adjusting its current summative model to assessments that are minimally disruptive to instructional time and form a progress monitoring system that provides timely data and information to support instruction. This cumulative scoring model takes into account student proficiency demonstrated throughout the

year and supports training for teachers and administrators on the interpretation and use of the data. In this model, the aim is for students to understand their progress, track toward grade-level proficiency, and have greater ownership over their learning. Other aims are for teachers to use TTAP data to identify students who need intervention and for administrators to use the information to support campuses and teachers better. The ultimate aim is to generate positive short- and long-term outcomes (e.g., a better testing experience for students). Because all students must take the current summative assessment (STAAR) until that system is replaced, Texas will be able to study the TTAP's reliability, validity, and comparability with STAAR.

[Session 1D Overall Validity slides 29–38]
Dr. Nash discussed overall validity in the context of DLM, which has met all peer review requirements for use as an accountability assessment. DLM's theory of action (reviewed previously) is part of a three-tiered approach to assessment validation. The theory of action defines the statements or claims that must be in place to achieve the goals of the system (which encompass the intended uses). The interpretive argument defines the propositions that must be evaluated to support each statement or claim in the theory of action. Validity studies are identified to evaluate each proposition in the interpretive argument. States should summarize the evidence for each statement in the theory of action and for each proposition underlying the statement. Dr. Nash stressed that this is a crucial component for peer review, as the accumulation of evidence makes the argument for the assessment system. States should further categorize the supporting information according to the five types of evidence for validity defined by the American Psychological Association and American Educational Research Association's standards: content, response process, internal consistency, relation to other variables, and consequences. She reviewed PIE, a CGSA-funded grant project, as an example.

[Session 1D Overall Validity slides 39–41]
Dr. Winter focused on responding to peer review requirements, noting that theories of action can have many uses apart from making the case for validity. For example, one state uses its theory of action to communicate the goals of the assessment system and progress made to the public. Theories of action can be used to explain trade-offs to policy professionals and demonstrate how proposed assessment features will affect outcomes. The key is to describe the system's purpose and use and to link them to evidence for the relevant CEs, define how scores are used, and clarify uses that are outside of peer review purview. States that use matrix sampling should show how student scores represent content domains as a whole. For example, they might present evidence of how student-level scores are represented in assessment design. Presenting procedural evidence with a literature-based rationale will facilitate successful peer review.

Questions and Comments

Does the validity evidence differ for TYA compared with traditional end-of-year assessment?

Dr. Nash replied that the evidence will not be very different for those approaches. One could argue that for TYA, states need to go into more depth on the content structures of assessment,

administration model, and infrastructure design to help peers understand this system. She added that states should explain what they are trying to accomplish with the new assessment system and rely on evidence that they already collect. Peer review submissions should cover every part of the assessment process. Sometimes the evidence will look different, such as when a scoring model other than IRT is used. In those cases, states should present an explanation. For the TYA approach, it will be important to describe monitoring and ways to address flexible design and implementation. Dr. Winter added that TYA has different goals, purposes, and outcomes than end-of-year assessment. For example, TYA offers instructional information in real time, as well as summative assessment. Dr. Warner suggested that when presenting a complex system of assessment, the theory of action could be discussed as a series of "if this, then that" regarding the evidence offered to describe what might happen. For example, with TYA, the timing of test administrations can vary and depends on many factors—and the state needs to explain this.

What has worked well for cognitive labs as supplements or alternative assessments?

Dr. Winter advised that states should consult with ED before making final decisions on cognitive labs. She stressed that a requirement for cognitive processes is not the same as a requirement for cognitive complexity and depends on the standard. The idea is to determine the intended cognitive process (what the child is thinking) and what the student intends to do—that is, how the student is understanding the task and approaches problem-solving. The assessment should address the task in a manner that reflects the standards and their operationalization in the state. Alternatives to cognitive labs include small class tryouts, in which the teacher administers prototype items to the class and discusses them with students. Documentation helps determine whether the cognitive processes intended are actually measured. Another option is to use an alignment process in which experts explain the cognitive processes required for the task. States should check with ED on whether this is acceptable. States might present research on other types of assessments that are similar and then validate their own tests. Cognitive labs offer good information on accessibility and the design process but are expensive. Dr. Dadey noted that expert judgment can be helpful, but states need to provide more evidence. He added that cognitive labs are wide-ranging and that research is needed to adapt them for specific contexts. ED offers information on cognitive labs on its website.

If score interpretation is affected by flexibility (e.g., design), how can that be made clear in reporting?

Ms. Rozunick commented that communication is key in this situation and stressed that states should plan for their explanations. She suggested talking to education professionals in pilot districts prior to developing such communications to determine the main points to address. Dr. Winter added that flexibilities that are part of accessibility or accommodations do not need to be reported. However, flexibilities that are part of the system should be communicated to the public and are usually received positively. Dr. Warner remarked that flexibilities are a common occurrence with some assessments—such as ELP—and the different modes of testing are

comparable by design. He suggested using ELP reporting as a model and ensuring that the validity of the scores comes through in reporting.