Session A-3: Validity Evidence: Content and Cognitive/Linguistic Processes
Panelists: ***Angela Broaddus, Ryan Kettler, Tracey Hembry, William Lorié***
Moderator: ***Susan Weigert, U.S. Department of Education, Office of Special Education and Rehabilitative Services***

In this session, panelists focused on two critical elements that impact the validity argument made by states in their peer review submissions (critical elements 3.1 and 3.2). Specifically, panelists addressed the content validity evidence that is effective, as well as evidence of cognitive processes (for academic assessments) and/or linguistic processes (for ELP assessments) that support a state's validity claims. Ms. Susan Weigert introduced the panelists.

Mr. Ryan Kettler led a general discussion on validity. Embedded in the Guide under 3.1 is a reference to overall validity, which Mr. Kettler interpreted as being connected to construct validity, which is the degree to which the scores evaluated represent the constructs that should be represented to yield accurate and appropriate inferences and interpretations. All these types of evidence connect to at least the most basic inferences, which is that the scores on reading tests reflect students' reading abilities, and the scores on math tests reflect students' math abilities.

The construct validity piece connects equally to critical elements 3.1 through 3.4., four different types of evidence that collectively build an argument for construct validity evidence. Highlighting what is new in this area, Mr. Kettler read the language from the Guide requirement: "The state has documented adequate overall validity evidence for its assessments consistent with nationally recognized professional and technical testing standards."

Mr. Kettler said the best known of these sets of standards are the Standards for Educational and Psychological Testing, which were revised several years ago by a joint committee of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement and Education. The standards in the four areas should be used in a thoughtful way to gather evidence to support the interpretations and uses of tests. The standards should not all have equal weights in each case. However, it is important to collect evidence in each of these four areas because each speaks to the most basic inference—that the reading score represents the reading construct and the math score represents math ability.

The Guide includes examples to help states meet the requirements of 3.1. For instance, states can have a validity chapter in a technical manual on assessments that addresses each of the four areas, at a minimum. The chapter should also, at a minimum, indicate the uses, the interpretations, and the intended inferences from a test. Building beyond the idea that all evidence is construct validity evidence is the idea that a validity argument should be built around those intended inferences, uses, and interpretations of the test.

This line of thinking is most often attributed to Michael Kane, who has written extensively in this area. Collectively, each basic type of evidence—content, response processes, internal structure validity evidence, and relations with other measures—feeds into the idea that the score represents what it is intended to represent.

There are inferences and uses that states often want to use that go beyond the most basic inference. Kane's point that validity arguments should be built around intended inferences would logically indicate,

for example, that to evaluate teachers, school buildings, or LEAs based on these scores, evidence needs to be collected not only at the student level, which is often the case for reliability and validity evidence, but also at the classroom and building levels.

If a growth score is being used, such as a student growth percentile or a conditional growth index to evaluate students, then evidence should be collected that not only indicates that the before and after scores are appropriate, but also that the growth index represents growth. Content, relations with other variables, and other types of evidence should connect to that growth score.

As another example, to make the argument that scores from a reading or math test inform intervention in some way, best practice requires collecting evidence indicating that persons who have lower scores on specific subscales or on the overall reading or mathematics score benefit, or perhaps even benefit disproportionately compared to other students, when they are connected to certain interventions.

To make inferences beyond the basic inference that reading represents reading and math represents math, evidence must be collected that indicates that these scores can be used appropriately for those purposes.

Mr. Steve Ferrara noted he had heard discussions about restructuring technical reports as validity item reports. Mr. Kettler said he had not heard about this issue and said states typically indicate intended uses, but were not following the strict protocol laid out by Michael Kane. Ms. Tracey Hembry stated she had not seen that issue as a peer reviewer.

Ms. Hembry continued the presentation with a closer look at three ways states often respond to critical element 3.1.

Alignment with State Standards

For academic assessments, states typically provide evidence through alignment studies. According to federal regulations, for every test, there should be an alignment study. This is typically done by a vendor, but not always, said Ms. Hembry, who has provided psychometric expertise in both education and credentialing for more than 10 years. When conducting an alignment study using the test blueprint, if standards are not part of the objectives, the alignment study should be going not just from the items to the objectives, but also from the objectives back to the standards.

States often submit evidence from a vendor. In the case of an alignment study, the vendor will develop a report for the state to submit. Ms. Hembry said that is all the evidence the state needs to submit. In alignment studies, however, items rarely align perfectly to the standards. Ms. Hembry recommended adding narrative in the index document about what comes next to show how the vendor's findings will inform the program, moving forward. This is the "now what?" question for which peer reviewers want answers.

Adherence to the Blueprint

In an adaptive assessment, peer reviewers see tables about the percentage of students who received a test that aligns with and adheres to the blueprint. As with an alignment study, peer reviewers want to see a narrative explanation if a state did not adhere to the blueprint for every student. Ms. Hembry said

the narrative should explain next steps for moving forward. The goal is not perfection; instead, states should provide available information and explain plans for improvement.

Robustness of Item Bank

Evidence is required that an item bank can support the blueprint and design of assessments. Again, if this does not adhere perfectly or is not as robust as it should be, states should document the development efforts they will undertake to meet those requirements.

Questions and Comments

- In response to a question from Ms. Jan Sheinker, Ms. Hembry noted that the vendor who conducts the alignment study is typically independent.
- Another question addressed the timing of alignment studies, i.e., whether these studies should be done before or after assembling forms. Ms. Hembry said peer reviewers do not have any set rules about the timing. Further, having worked with states as a vendor, Ms. Hembry said the timing could depend on the schedule and other moving parts in the process.
- Ms. Hembry also addressed off-grade-level testing. She said the Guide requires that proficiency levels be determined based on grade-level content. Once that determination is made, items can be administered beyond that if a student performs at a lower or higher level.
- To clarify an earlier comment, Ms. Hembry said that having an alignment study for every test for the academic assessments means every grade level and all content.

Ms. Angela Broaddus continued the discussion of critical element 3.1, noting additional considerations for alternate assessments.

The states should meet the requirement that the AAAS are aligned to the adopted standards for the state. That can be done through a set of extended standards. That helps practitioners but creates an additional validity step because all the links need to be clear to a reviewer. Ms. Broaddus noted that she has worked for several years as a mathematics content specialist on the development of general, alternate, and formative assessments used in multiple states.

Another consideration for alternate assessments is alignment at grade level. States should ensure that the items students receive are aligned to the appropriate grade level, meaning the assigned grade level for that student. That argument must be made clear to the reviewers.

Another consideration Ms. Hembry noted was that for adaptive tests, reviewers generally seek to understand how each test adheres to the breadth and depth of the standards. For alternate assessments, the issues are breadth and reduced complexity. Reviewers seek information about how the state describes the alternate assessment in terms of ensuring breadth and how the state has reduced complexity to be appropriate for students in the population.

Addressing 3.2, Ms. Broaddus noted the challenge of trying to provide evidence for this critical element. Cognitive labs remain the gold standard, but they are expensive and difficult for numerous reasons. As a peer reviewer, she has seen cognitive labs that have been very brief, meaning they had few students and few items. There was no evidence of the breadth of the assessment programming having been studied in this way. States should supply evidence describing how a significant number of items operate

with a strong sample of students from different grades and content areas. Other forms of evidence can include the following:

- Think aloud studies;
- Analyses of incorrect response options (sometimes called distractor analyses);
- Surveys of students after tests;
- Time spent on items; and
- Skipping patterns.

These options are available in different ways for online testing.

A good discussion by the state indicating how this information will inform the assessment program remains very important to the review process. If items are drawing on the intended cognitive processes, perhaps that item template gets populated more. If not, perhaps a state will make adjustments. In any case, states should include a discussion of how the findings of investigations into cognitive processes will inform test development.

Ms. Broaddus stated that the evidence-centered design process recommends item templates or families of items that hang together and are proposed to do the same thing. State or programs could study an item template or one or two samples from that item template to avoid overstudying one type of item. That might be a way to reduce costs while getting more "bang for the buck."
The Guide also notes these sources of cognitive process evidence:

- **Expert judgment.** States should consider supplying evidence that describes committee membership, including both demographics and the members' expertise. A description of panel membership for these types of reviews is helpful.
- **Relationships between responses and other useful variables.** To assess a student's ability to think mathematically, a state could determine whether teacher reports say the same thing. Looking at relationships among variables is another way to describe the cognitive processes that are evoked by certain families of items. States should supply evidence that shows how the findings of such studies inform test development.

Questions and Comments

Mr. Anton Jackson asked about other schemes for the Next Generation of Science Standards that do not fit well into cognitive models. Ms. Broaddus noted depth of knowledge.

Mr. Peasley noted that in the Guide, ED does not endorse a particular alignment model or approach because there are multiple acceptable approaches to assembling this evidence. Further, peer reviewers sort the submission to see if the document offers a reasonable set of evidence, given the requirements.

Mr. Gary Cook, who serves as a peer reviewer, said he has a personal opinion about alignment, but considers a state's reasonable methodology and set of arguments. States should have a reasoned approach that looks at depth, breadth, and reasoned methodology. Ms. Hembry added that psychometricians might have preferred methods, but peers do not expect states to submit information in a specific way. States should choose what works for them and explain why.

Mr. Harold Doran of the American Institutes for Research (AIR) said some of the studies are easy to conduct with data. Cognitive labs or consequential validity are more difficult. A state could invest millions in cognitive labs and then receive negative comments from peer reviewers. Perhaps states should submit a proposed design, receive feedback from the peer reviewers, and if the study is done with the fidelity described, it would receive approval. Ms. Broaddus said peer reviewers acknowledge the difficulties of providing evidence. The purpose of the session was simply to provide suggestions.

Ms. Broaddus commented on teacher reports and referred to a paragraph in the Guide on other measures that require similar levels of cognitive complexity in the content area, such as teacher ratings of students' performance. A teacher or a test developer could give students both multiple-choice questions and open-ended responses and look at the results together.

Session participants requested examples of a good cognitive lab report. Mr. Kettler agreed that this is some of the costliest evidence to obtain. He said a good cognitive lab doesn't require a large number of participants. A representative sample across grade levels, disabilities statuses, and so on, could suffice. Educators also could partner with local universities where education psychology or special education students are looking for data to complete theses or dissertations. Further, surveys after tests that ask students for feedback on various questions could add value. Distractor analyses, recording the amount of time, or teacher ratings on how a person would have to respond are less costly forms of data to obtain. They can fit in element 3.2 if a state has a nuanced argument for showing how they connect to validity evidence based on response processes.

Mr. William Lorié said one of the most difficult aspects of 3.2 is thinking about the element critically and thoughtfully. Cognitive processes refer to something very specific. Peer reviewers want to see where states think cognitive processes are applicable, what those cognitive processes are, and how they are being investigated in a focused manner.

Mr. Lorié noted two approaches to cognitive processes that have been used in the best technical reports he has seen in this area. He said both need to be addressed in some programs. One is the sense of cognitive processes that is some sort reasoning (i.e., that the student is reasoning or applying knowledge or skills in a certain way to respond to a type of question). Peer reviewers want evidence that students have been reasoning or following a certain process for a type of item because that reasoning or that process is implied in the standard. The second approach to response processes in the standards is the format of an item. If the format of an item is new or atypical, peer reviewers want evidence that the students understand the instructions for the item, can internalize the instructions, and are not responding to the item in a way that is inconsistent with those instructions. Mr. Lorié provided four useful steps for addressing cognitive standards:

1. Survey the standards to see where cognitive processes are implied.
2. Identify the cognitive process. What is the cognitive process or processes for which a state is seeking evidence?
3. What observable student actions or statements would provide that evidence? Perhaps a state could ask questions right after the students respond to an item or set of items.
4. How will a state create the conditions to elicit those actions or statements if the students are indeed following the cognitive processes implied in the standards?

Session participants asked questions about determining cognitive processes. Mr. Lorié said if the standard is about an outcome or a product that the student can do, and it is not addressing the manner

in which the outcome occurs or a specific cognitive process in the standards, then cognitive process would not be relevant.

In thinking about cognitive processes, states should consider whether accommodations erode the intended cognitive processes of a set of items, said Ms. Broaddus. Further, is the cognitive process eroded beyond the intention of the accommodation?

Ms. Broaddus also addressed alternate assessments with regard to cognitive processes.
The standards for reporting whether the intended cognitive processes are evoked remain the same, but collecting evidence from that population can be difficult. Instead of having these students think aloud while completing a performance task, Mr. Kettler recommended getting feedback from the students immediately afterwards. Teacher ratings also could assist in the argument.

Mr. Kettler discussed critical elements 3.3 and 3.4. Internal structure validity evidence addresses whether the various pieces of a test fit and work together in a way that would be predicted by the underlying theory of the test. States should think about how those numbers and operations items, geometry items, and algebra items fit together to contribute to a mathematics score.
He commented on what has worked well, saying that regardless of the use of exploratory factor analysis, confirmatory factor analysis, or the correlations among subscores, the underlying theory on how the different parts of a test fit together should match how they actually fit together. For instance, does each subscale contribute to the overall score or not?

If an underlying theory doesn't match the evidence or doesn't match the reporting structure at the end, peer reviewers might become confused.

Mr. Kettler said that the relationship between critical element 3.4 and other variables shows how well scores (whether a correlation or agreement indices) relate to the other scores to which they are supposed to relate. Peer reviewers hope that the requirements of the basic multitrait/multimethod matrix are met. For instance, if peer reviewers see higher relationships among interim math and reading scores than between measures for math or reading, it's problematic because it shows that scores are driven more by the method and time of testing than by the construct being measured.

Mr. Doran noted that peer reviewers should remember that certain models are used because that is how it has always been done, and unless subscale scores are reported, state leaders may not be satisfied. Sometimes certain decisions are due to a state's political complexities. Mr. Ferrara noted that states could include a comment in the peer review submission about the practical need for reporting subscores, even though an argument was made on a unidimensional model. Ms. Hembry said peer reviewers appreciate hearing the rationale from the states in the index document.

Ms. Shelley Loving Ryder discussed the effect of accommodations on Virginia's reading test. A study looked at allowing a read-aloud for students with disabilities, and how that compared to students who did not have the read-aloud. The state looked at the mean scale scores of both groups and found very little difference. That justified providing the read-aloud for very specific students who have an identified issue, either with decoding or a visual impairment, and have not yet learned Braille.