Session D-2: Validity Evidence: Content and Cognitive/Linguistic Processes
Panelists: ***Jeffrey Hauger, June Zack, Ryan Kettler, William Lorié***
Moderator: ***Jessica McKinney, U.S. Department of Education, Office of State Support***

Panelists focused this session on two critical elements that impact the validity argument made by states in their peer review submissions. Specifically, panelists addressed effective content validity evidence, as well as evidence of cognitive process (for academic assessments) and/or linguistic process (ELP assessments) that support a state's validity claims. This session addressed critical elements 3.1 and 3.2. Mr. Ryan Kettler described overall content validity, which is the degree to which a score in a specific area represents the construct (i.e., the math score represents how the student is doing in mathematics). Each type of evidence should not be equally weighted; the four different types should be looked at together. Additionally, he noted the following:

- The best-known standards are the Standards for Educational and Psychological Testing. Michael Kane has written extensively on evidence validity arguments. Building interpretation and uses can guide what the states do.
- Reliability and validity should be collected at the aggregate level if states want to make inferences about mean scores.
- If your intended uses are connected to intervention programming, students who are low in certain areas benefit from the interventions they are connected to. Evidence of benefits is costly to gather.

Mr. Jeffrey Hauger addressed content validity, stating that you can't just look at the breadth; you have to look at the depth as well. States should have an independent source verify validity, rather than just the vendor. It needs to be consistent with nationally recognized standards. Clearly address in the notes how your data fit in with the overall validity standards. Be direct, and use the comments to build the story of the validity evidence.

Mr. William Lorié stated that for 3.2, cognitive labs explain what their program is going to do with the findings. Although cognitive labs are the gold standard, another method specified in the evidence for 3.2 is expert judgment. He suggested look at the composition of the expert panel and their process. Do they list the processes? How do they make their judgments?

Mr. Lorié said it can be a challenge for submitters to think critically. He noted that the academic standards part has not changed in the new Guide. The challenge is to look at the testing program in terms of cognitive processes. In some cases, cognitive processes are not relevant. States need to focus on where to look for evidence of cognitive processes. What actions can be observed to provide the evidence needed? Sometimes you can ask students after they've taken sample items. Collect data early in the test development process. How will you create conditions to elicit those actions? The student uses the context for the word or phrase to make a determination of what it means. Be very focused in your search for cognitive processes.

Ms. June Zack said validity must be embedded at the beginning, and states have to plan for it and use an evidence-centered design. It starts when you develop your blueprint and conduct your item reviews. Peers have to review for the cognitive and linguistic processes you are purporting to measure. You should have review logs and collect data from experts. Ask: Does it cover all standards and domains? Pick a model that fits your design, follow it through, and refer to it at each step of the way as you're

collecting validity evidence. Ms. Zack recommended cognitive labs. She said to take items that you intend to investigate, but make sure your student sample represents your testing population. Decide what evidence you want, and plan for your data collection instrument to collect it. Then run the cognitive lab. She conducts cognitive labs inexpensively by having state personnel run them. Alternatively, teachers can be trained to run cognitive labs to lower costs. You get information that is instrumental in designing your items. Then you are on your way to having a solid assessment because your item pool is tapping what you want to assess. Seek other independent expert judgments with a research company or vendor other than the one you use for your assessment. There are relationships with other measures; it's hard to do, especially with an alternate population. There aren't other measures available that tap the same thing you're testing with your standards. Not every LEA uses the same measures to determine who should be in special education and who should take the alternate assessment. It is hard to find measures that are used consistently by LEAs within a state to correlate data and see if you have a match.

Mr. Kettler noted critical element 3.3 on internal structure. What works well is any alliance between the theory the test is based on, the relationship between the parts of the test, and the reporting that's done at the end. What doesn't work well is when the theory doesn't match the other aspects.

Critical element 3.4, relationships to other variables, must show that scores relate to other variables in the magnitudes expected. You want to see correlations at higher magnitudes than those within method. He said he likes a multiframe, multimethod framework.

Mr. Lorié added that response processes cover cognitive processes, but there are unusual item formats that can get in the way of the construct. They may not understand the response processes. This falls into the 3.2 arena.

Ms. Zack said cognitive labs can help you determine whether different items really work, whether you are getting the information you want, and whether students are using the cognitive processes you want them to use. She suggested running a small or informal cognitive lab to determine if item formats are working, and said sometimes a plain multiple-choice item is better than something "new" or "jazzy."

Questions and Comments

- Mr. Lorié said many submissions are developed with the assumption that a unidimensional scale has to be used, but this is a hard requirement to meet. It doesn't have to be the model for testing programs that want accurate subscores. There other options for obtaining reliable and independent scores other than a unidimensional scale.
- Another participant asked if it's as simple as asking, "What is the cognitive level, and does the item tap the cognitive level (yes or no), or should there be more on the review sheet?" Ms. Zack said asking these questions is not the same as having a cognitive lab. You can ask, "What do you think the student has to do to answer this question?" She said teachers don't discuss the cognitive process used enough. Ask them what cognitive process they will use, whether it is valid, and whether it aligns with what you want the student to do. You want to confirm whether the teacher's expectations are met in peer review. It can be helpful to survey students to ask "What were you thinking about?" A distractor analysis can be used. This can be valuable if the distractors really test different things and you plan them well. Distractor analysis works well with math and science items, but is hard to do with English language arts and social studies. On

computer-based tests, the amount of time a student spends on an item can be recorded. A cognitive lab doesn't have to have a large sample, but it should be representative in terms of academic background. Be open to the possibility that graduate students at a local university might want to work on a cognitive lab for a thesis or dissertation.

- A participant whose state has a centrally scored data portfolio asked, "What would you look at?" Ms. Zack said to formulate what you're going to collect ahead of time and what standards you are assessing. You can have a committee look at your plan ahead of time. You can include interim products in the portfolio, not just the final products and the standards the portfolio is addressing, such as a writing standard.
- Ms. Jessica McKinney asked what a state should do when areas of weakness are found. Panelists agreed that states should provide information in their submissions stating what they intend to do about these problems. Explain when you will achieve your end goal. Mr. Kettler added that states need to be proactive and explain what areas they're still working on. Communication is important; if they don't tell the peer reviewers about weaknesses, the reviewers might assume you believe all your evidence is good. Almost everyone has an alignment report with weaknesses; the best case is to explain that at the end of the timeline, you will have a revised test or have your changes independently reviewed.
- A participant asked about strategies for dealing with resistance to independent alignment studies. Ms. Zack said she would be surprised if a vendor blocked them; independent alignment studies only help. If you are confident in the process you used, why would the findings upset you? If they find something deficient, then you have an issue you can to attend to. A bigger problem for states is cost, which can prevent these studies.
- A participant asked, "What are some key considerations for items to include in a cognitive lab?" The panelists said not to start with items; look at the standards that apply to cognitive processes. Use a broad sample of representative students; it doesn't have to be large. Sometimes states have a small, closely guarded item pool. However, once you use items for a cognitive lab, those items are no longer secure. States don't want to put them on their summary assessments. To address concerns about the item pool, you might want a separate sample. Be thoughtful about the items on the practice test; don't delete your operational pool. Also, it's more expensive if you don't plan ahead. The story of how the state collects the validity evidence is really important.
- The closing question asked the panel to describe the biggest problems for peer reviewers. They said it's when they receive a "data dump" of multiple large manuals that are not labeled, and the state expects the peers to go through them. That's why states need to tell a story and make an argument, rather than provide extensive unorganized information.