Session C-3: Validity Evidence for English Language Proficiency (ELP) Assessments
Moderator: ***Brenda Calderon, U.S. Department of Education, Office of State Support***
Panelists: ***Gary Cook, Edynn Sato, Phoebe Winter***

Panelists discussed the requirements in the updated Guide as they relate to ELP assessment validity and alternate ELP assessment (AELPA) peer review (critical elements 2.1 and 2.2, and section 3). Particular attention was paid to overall test design and item development, and to the four types of validity evidence found in the Guide. Panelists discussed practical interpretations of the ESSA requirement that ELP standards be aligned with the state's academic standards.

Ms. Brenda Calderon opened the session. Ms. Edynn Sato led the presentation with a focus on key issues in alternate ELP assessment. She said this is a fairly new area in terms of the requirements and standards that states should meet. Although alternate ELPs have been in existence, states must look at the data differently to ensure that there is solid evidence to meet the critical elements in the Guide.

Critical challenges related to validity include the following:

- **Population:** Data were shared in previous breakout sessions on how states can look at this population and gain more confidence in understanding and identifying students properly. States have experienced both over- and under-identification. There is no federally designated category, so states must make an effort to understand students with significant cognitive disabilities. Further, for these students also are ELs, states must identify language and disability needs. Until states understand these students, designing an assessment for this population and eliciting what students know and can do in terms of the English language will remain a challenge and a fundamental validity issue. This information has implications for the design of assessment items, the test blueprint, participation guidelines, accommodations, and more.
- **Construct:** Defining the construct ties back to understanding the population and its characteristics. Understanding how these students demonstrate English language proficiency in listening, speaking, reading, and writing is key. States also must distinguish between English language proficiency, knowledge and skills, and disability. States should focus on English language proficiency development rather than disability needs, said Ms. Sato. States also must define operationalizing listening, speaking, reading, and writing in a range of ways that is fair and reasonable for these students, given how they typically demonstrate what they know and can do across the domains.
- **Language acquisition:** According to the requirements, students must demonstrate proficiency in the English language to access and achieve grade-level curricula. States must be clear on the differences between communication versus gaining English language proficiency more broadly (being able to speak in English in a general way) versus English language knowledge and skills that enable students to access, engage with, and achieve academic curricula. Ms. Sato highlighted these nuanced differences and encouraged states to develop clear definitions within a construct and then focus on operationalizing those definitions. This is a fundamental validity issue.
- **Theory of action:** A state can use a theory of action to track the key pieces that underlie the validity argument for its assessment.
- **K-2 population:** These students are not tested with alternate academic assessments, so states must develop a process for correctly identifying this population and providing tools to help

students make progress toward proficiency. The states are responsible for identifying these students and tracking progress toward proficiency and achievement.

Available resources to help states define and understand this population include the ALTELLA, an Enhanced Assessment Instruments Grants-based project. The WIDA Consortium has had alternate ACCESS in place for years, as well as data to support available processes and reports. ELPA21 has a white paper with definitions and processes for identifying students, participation guidelines, and a theory of action. The National Center on Educational Outcomes also has a number of white papers. CCSSO continues to develop a resource related to ELs with significant cognitive disabilities and alternate ELP standards for K-12.

For construct issues, many of these organizations also have resources available on processes, how to define domains, and what English language proficiency means for this population of students. WIDA has a process for identifying students in K-2. A white paper by Patricia Almond and others (based on discussions with SRI International in 2011) offers guidance on cognitive labs related to students with significant cognitive disabilities.

Ms. Phoebe Winter discussed validity issues related to AELPA development. States should support validity with a body of evidence, and that body of evidence should begin with test and item design and development. Ms. Winter's work with state and nongovernmental education agencies focuses on bringing policy, psychometric, and practical perspectives to the design and implementation of educational assessment and accountability programs.

Focusing primarily on sections 2.1 and 2.2, Ms. Winter said the underlying theory of second language acquisition should be explicit. Further, a state's theory of action or claims about test outcomes should tie into this underlying theory. The theory will influence how states address the link to language required to access and demonstrate skills in academic content. It also will determine whether a state structures a test around tasks or individual items, whether each standard receives equal weight in each grade band, and other issues. The underlying theory of second language acquisition will affect everything, including how a state defines alignment.

A theory of action also will affect how a state determines whether a test reflects the depth and breadth of its standards. For example, a state might weigh a standard more heavily in the early grades than in the later grades. Does the state need equal weight in all grades, given that some content alignment rules will not necessarily apply to an English language proficiency assessment? A theory-based rationale will help a state develop its test.

Ms. Winter noted that the development aspects in section 2, the validity aspects in section 3, and the technical quality aspects in section 4 will intermingle. If states discuss alignment in 2.1, the topic does not have to be readdressed in 3.1. States should show evidence for alignment where it fits in with test design, then point back to it in 3.1.

Discussing test design and development further, Ms. Winter made the following points:

- States should show how the underlying theory manifests in the blueprint and test design. They should connect the blueprint and test design to the theory of language acquisition. In section 2.2, states can point back to 2.1 to acknowledge how the critical element was met.

- As with content assessments, states that use evidence-centered design or another item design and test development process can easily lay out the connection between standards and assessment. Evidence-centered design and other strategies lead peer reviewers along the path. The Guide is structured on the idea of these connections.
- States should document how conceptions of accessibility and accommodations were included in the test design itself. Accommodation is not separate and off to the side. States must consider the whole population of students when designing the test and items. For instance, states might not include certain items because no appropriate accommodations are available. Or states might include an item and decide to include parallel items for certain disabilities. For students who use Braille, states might design two items at the same time: one that can be put into Braille and one that cannot. States must discuss those efforts in section 2.2 and point back to that information in other sections of the guidelines.

States also should show consideration for the following three aspects of English language proficiency tests during test design:

- The basis of the test content and the four domains of language. The domains must be evident in state standards.
- The full range of linguistic complexity as reflected in proficiency standards.
- The connection to the language needed to acquire the knowledge and skills in the academic content standards.

State test design must reflect all three areas, said Ms. Winter. This information might already be available in a technical manual and in foundational documents such as the theory of action and the validity agenda.

States can discuss alignment in several places within the evidence requirement. For content assessment, states should fully describe an independent alignment review that requires participants to consider content match, linguistic complexity, and connection to academic language in at least reading, math, and science.

Ms. Winter said that for item development, many of the processes are the same for academic assessment. States should follow good practice in developing test items, remembering the key points of linguistic complexity and connection to academic content standards.

States can conduct cognitive labs before or after administering a test. When conducting an English language proficiency test, someone must be available who has the same language background as the students to collect information on what students are thinking. States can select items that illustrate the types of items on an assessment or items that raise concerns. States that are unsure how a group of students will react to a certain test item should include a few of those students in the cognitive lab. These labs should be conducted purposefully and carefully.

An alignment study does not provide the kind of information a state needs about eliciting intended linguistic processes. This type of study provides some content and depth, but states should ask deeper questions. Another way for states to get good data is through small-group discussions on what students did to respond to a set of items. States should think of creative ways to get information about linguistic processes.

Mr. Cook highlighted validity issues related to analysis. He noted that peer reviewers often see cognitive labs that are post hoc. They see alignment studies that have been done, but the follow-up submission does not show what the state plans to do about the findings. For instance, how will test design change as a result? Simply providing materials about a study or a cognitive lab is insufficient.

For internal structure and relationship to other variables, peer reviewers look for an argument to show whether the internal structure of the test reflects the construct the state claims to measure. At a minimum, one might expect something to associate with listening, reading, speaking, and writing, and some sort of structural analysis that would demonstrate that. Communicate an underlying structure associated with the test that reflects the domains, that differentiates the levels, and in some form or fashion, associates with the topic area. Mr. Cook did not want to suggest a specific way to operationalize this effort because states can use a variety of strategies in a peer review process. Evidence in a technical manual or independent studies can support that information, and both ELPA21 and WIDA can assist in the process. The goal is to show how the test scores reported represent the underlying construct claimed.

Federal law defines English learners as students whose language proficiency prevents them from being proficient in a state academic content test, prevents them from meaningfully interacting in the classroom where the language instruction is only in English, and prevents them from meaningfully participating in society. Mr. Cook made the following points associated with that definition that are relevant to the idea of external variables:

- **State academic assessments:** Mr. Cook has used this variable to ask such questions as "What does proficiency look like?"
- **Classroom performance:** Peer reviewers would expect students that are reported by teachers as being at a high level in the classroom to be represented in assessments as high-performing students. Likewise, if a student is low-performing in the classroom, peer reviewers would expect low performance on assessments. Peer reviewers should be able to look at the classroom relationships between assessments and how teachers perceive that students are performing.

Questions and Comments

Mr. Dan Wiener from Massachusetts asked about how WIDA states have diverged in the development of exit criteria. He asked if that was tied to one of the elements where states have different criteria used to exit students from EL status. Mr. Cook said the idea of exit criteria has to do with accountability and reclassification, and that differs across states. WIDA must show a full range of language proficiency, sufficient to help states make decisions about students at various proficiency levels.

Ms. Kristine David asked about the number of participants needed in a cognitive lab for the alternate population. Ms. Winter suggested looking for a rationale for the number of participants the state already has. If a state wants to make sure that students with different disability characteristics are able to access a test, it should have enough students at enough grade levels representing enough disability characteristics to answer the question. The answer also depends on what a state is looking for. For something very specific, a state might need fewer students in a cognitive lab than a state looking at the more general aspects of a test.

Ms. Sato agreed, noting that the specific question a state wants to address drives the number of participants. Ms. Winter reminded the audience that cognitive labs are only a start. These efforts help

states understand the kinds of questions to ask in a larger-scale analysis. The panel also encouraged states to consider tryouts with large groups of students and to rely on a TAC for guidance.