Plenary Session Two: Focus on ELP Assessment Peer Review: General ELP Assessments and Alternate ELP Assessments
Moderator: ***Amy Bae, U.S. Department of Education, Office of Special Education and Rehabilitative Services***
Panel One*:* General ELP Assessments
Panelists: ***Eric Zilbert, Gary Cook, Margaret Ho, Phoebe Winter***

A significant new requirements in ESSA is that ELP assessments are now part of the state assessment peer review process. As this is new for all states, two panels discussed (1) considerations for general ELP assessment peer review; and (2) considerations for alternate ELP peer review. Panelists used examples and content from the updated Guide to frame the discussion and to guide questions.

The moderator, Ms. Amy Bae, stated that there is a notable change in ESSA that requires ELP assessments as part of the peer review process. Ms. Phoebe Winter noted that Section 1 addresses statewide systems of assessment. The requirements are very different for ELP versus academic content. Standards must (1) be derived from the four recognized domains of speaking, listening, reading, and writing; (2) address the different proficiency levels of ELs; and (3) align or correspond with challenging state academic standards. Proficiency standards are often written first and may not correspond to cut scores. ELP standards must align with the language demands of reading/language arts, mathematics, and science. Evidence of alignment can be demonstrated (1) by a strong correspondence between state and ELP standards, (2) by adequately demonstrating specific skills or (3) by submitting reports of an external review of a state's ELP standards. The alignment requirement isn't new, but documentation for peer review purposes has received little attention; therefore, heavily researched models do not exist, although there are some strong models on alternate achievement models from content area to content area, and some evidence of alignment can be found in the documentation of the development of the standards.

Mr. Gary Cook noted that there has been no rigorous organization of the available knowledge on alignment, as various methodologies haven't been looked at across countries, and they haven't been consolidated or submitted to an external peer review process. There also is no documented process. For a consortium, states will need help understanding the standards they are adopting.

Ms. Winter called attention to the phrase "including all students," which means including students with disabilities (critical elements 5.2 and 5.3). The requirements are the same for academic content and ELP in terms of making sure all students are able to take tests. However, the actual tools will differ. There is an issue because students with disabilities have had access to different tools, but they won't get them anymore. States need to prepare students for the different testing conditions they will have. She noted that changes in accommodations can't interfere with the communication constructs. States will need to rely on the theory underlying their standards to determine the acceptable accommodations. The literature is not strong on accommodations for ELP assessments for students with disabilities. Some things are translatable, but states will have to think about relying on best practices until there is more research.

Under critical element 5.1, the major difference is that a student can be exempted from being tested on items in a domain if there aren't appropriate accommodations for the domain, but scores must be provided based on the domains on which they are tested. Students who are deaf and/or blind will be most affected. Although students can be exempted, we should be interested in whether a student has

sufficient EL skills to acquire content in the classroom. They may sign in English or use a different system. We will need appropriate interpretations of the scores and what they tell us about the student.

Mr. Cook referenced critical item 2.1 and said that behind the standards, there are several different levels. For example, language proficiency has three dimensions: domain, level (more than one), and topic (must relate to the academic topic). There are several arguments about how to identify the underlying construct and how it includes all three of those dimensions. A unique aspect of this area of assessment, different than for content, is that the results may provide or deny students a constitutional right, in that identifying students a certain way could deny them services. Mr. Cook asked how the breadth and depth of the dimensions should be identified and suggested that the following should be considered during the item development process: how all three dimensions of language proficiency are resident, item review, and calibration that represents the dimensions.

Ms. Margaret Ho addressed the issue of how states can address the requirement for developing and selecting items in accordance with critical element 2.2. States must assess proficiency in terms of content and language processes. She noted CCSSO's 2012 English Language Proficiency Development (ELPD) Framework, which corresponds to the Common Core State Standards (CCSS). This resource (hereafter referred to as the Framework) may be useful in determining what correspondence means, as it outlines the underlying English language practices found in the CCSS and Next Generation Science Standards (NGSS), communicates the language that all ELs must acquire to successfully engage the CCSS and NGSS, and specifies a procedure by which to evaluate the degree of alignment between the Framework and ELP standards under consideration by states. Ms. Ho suggested looking at the ELP standards to see the extent to which those are reflected. This is the first important piece for item development. She suggested next looking at task types and blueprints. Those specifications are not very different than what the states are familiar with in terms of the underlying construct of what correspondence means.

Ms. Ho also addressed critical item 3.3, validity based on internal structure. There are four typical validity emphases that might be in a peer review narrative. For ELP, the intended language processes should align with the state's standards. Test forms are designed to cover multiple grades; they are not grade-specific. However, the expectation in terms of performance-level descriptors are state-specific and should relate to the content demands of a particular grade. She referenced Mr. Cook's language acquisition theory and noted that an assessment should incorporate a theoretical construct. She asked, "If a state has a cognitivist model, how does the state's test represent its language acquisition theory?" She noted that is a different component of the ELP assessment, and said those statements are typically in a state's theory of action. In the CCSSO Framework document, states can read about two approaches: the Formative Language Assessment Records (*FLARE*) Model and the functional model (what you do with language). She suggested exploring the work of Mr. Kenji Hakuta.

Mr. Eric Zilbert spoke about the link between scoring and reporting within critical items 4 and 6, and the necessity of "building an argument" for the link between the student's test score, the standards, and the score report. In the scoring area, specific to these tests, the four subscales must be measured. Most states already have a four-part test. If you're in a consortium, your lead will do that. The English Language Proficiency Assessments for California (ELPAC) were built from scratch. California carefully considered scoring and reporting. The technical report must articulate how you combined or estimated the subscale and came up with an overall score. The ELPAC will have an oral, listening/speaking scale and a reading/writing scale. It brings up the problem of high aspirations for what a test can do when the test must be taken in a reasonable period of time. This affects the services a student might get. They

have to make sure, through achievement-level descriptors, that they that link back to the standard. Everything must be documented for these assessments, just as for other assessments. In the speaking domain, scoring must be done on-site by teachers or testers in the LEA and scored locally. This requires developing training materials, training trainers and administrators on how to score, evaluating the training, submitting evaluations of training, and keeping rosters to show that trainings were held. Monitors need to make sure the scoring for speaking is done correctly. This requires having good documentation to share with peer reviewers on who helped develop achievement-level descriptors. California reports scores to parents and LEAs, include the scores in technical reports, and provides interpretive materials (to be provided in 10 languages). For constructed response items, Mr. Zilbert suggested using checks similar to those used for essays (that is, a method of back-reading).

For those that have adopted speaking/listening, language professors have been testing for many years, so they have been assessing ELP. However, states that report the inter- and intra-rater agreement of those who scored locally need to identify whether the raters are scoring reliably. Peers are almost certain to look at this. Local scoring states should monitor and check the reliability of the speaking tests.

Mr. Zilbert's state found standard setting to be complicated. Because of the importance of these cuts, which may refer a student for reclassification, the state does extra work to ensure that confidence in the cuts. He combines data and looks at how they compare. He said there is not one right answer about how much is enough. Theoreticians say you never finish being an El, that there is emerging, expanding, and bridging, but no arriving. However, we have to "arrive" to come up with reclassification. A comparative group study will look at the teachers' judgments and compare with panelists. Decision making is required on issues such as "Is level 2 good enough? Where is the sweet spot?" States will be submitting on the summative only. There is a section where they need to show a relationship between the screener and the summative, but the screener is not submitted.

The biggest challenges were monitoring, coordination among states using the same assessment, and the requirement to tie to an underlying theory of language acquisition.

Mr. Zilbert said states should read what's in the Guide and look closely at the examples when preparing for peer review. Then they should direct the peer reviewer to the evidence. Mr. Cook said to acknowledge when requirements aren't met and inform peer reviewers of the plan to remedy the situation. If there are differences in the standards and what you report, articulate why you deviated from the standards.

Panel Two: Alternate ELP Assessments
Panelists: *Audra Ahumada, Edynn Sato, Kim Brannan, Melissa Gholson*

This panel discussed considerations for the alternate ELP peer review. Ms. Melissa Gholson said that everything that was said about the general assessments applied to alternate ELP assessments, but there are additional requirements. She said strong submissions are always coherent. A practical approach is to begin by asking "What are we doing to identify the population of ELs with a significant cognitive disability?" This will lead you to think about test design and development. What are the characteristics of this population? It is a diverse population, and there will be domain considerations. How you allow these students to perform reading, writing, speaking, and listening will be very different than in a general assessment. Are there domains they can't perform in? Are some students deaf or blind? What does that mean for test design and scoring? The state will need a statement of the purpose of the assessment. On their project, they observed almost 100 ELs with significant cognitive disabilities. What are you doing to get to know the population? How are they being instructed? There are needs in the field; they want to inform teachers so they know what to do with the assessment results. It is a vulnerable population, and it's important that they are not exited too soon, yet they will also need to be ready to transition to the world of work. How do you define language complexity? If you are going to eliminate someone, how does that impact the composite score?

The moderator asked what the states to identify needs and challenges. Edynn Sato said she wanted to focus on three things: (1) population definition, (2) construct definition, and (3) accommodations. What do states need to do across the elements? There is no federal designation, so be clear on who these students are. That drives how you design items, and defines the construct. For population definition, the data you have used and who you have involved are some of the pieces of evidence to submit. For construct definition, some students have devices that help them. Given your understanding and definition, one of the next steps would be to determine what reading/writing/speaking look like for them. For accommodations, various levels of support are needed. Will the construct need to change? Think through what the accommodations look like and how that will affect the construct being measured.

Ms. Audra Ahumada asked about the challenges related to getting the best sets of data. She has seen some of the development of alternate ELP assessments as part of the National Center and State Collaborative (NCSC) Consortia. They used the Learner Characteristics Inventory (LCI). Based on work where students were being served, the Individual Characteristics Questionnaire (ICQ) was developed, which is similar to the LCI, and it helped them learn about these students. She noted that this was all teacher reported. They then developed task templates and test items. They used the LCI and ICQ to further develop assessment processes.

She said they ask the LEAs to report those who are designated as EL students. There is a large disconnect between what's being reported at the LEA level and what is reported in the surveys. There are many myths. Because these students don't communicate in the same ways as other students, they are not considered ELs. They need to indicate what is reported by teachers versus what is in the surveys. The teachers have longer lists. They are using multiple sets of data to look at this and determine who these students are.

Ms. Kim Brannan said they were in the early stages of this work in Texas. A year ago, they started looking at three high schools to get a basis. They looked at alternate assessment students and identified

about 4,500 in disability categories. They looked at how they were participating in the AELPA. Half were exempt in all domains. They did a deep dive, asking how they would assess these students. There are laws in Texas for this population; teachers can't make their own assessments. They got educators and stakeholders involved at all levels and started the conversation. They started a pilot test in a cognitive lab in the spring. They learned that many teachers were not bilingual and/or EL certified. Many felt the students didn't need EL services because they were getting special education services. Ms. Brannan said they need to get buy-in and have people believe in what they're doing. They don't have the answers, just information on what is working so far in Texas. They want to continuously improve. The conversation has been started, and the EL and special education staff are now talking to each other and working as a team. They just got data back on their items. They want educators to help them write about real classroom behaviors to make sure the documentation is authentic.

The moderator opened the floor to all panelists. One said that as they move forward with alternate ELP assessments, having grade-level expectations for these students is critical to the assessments. They should see a progression in expectations. There should be a range of allowable performance. Some might know English but are designated as ELs. Conversely, some students don't receive services, so how will they learn English? How will they gain college and career readiness skills? She cautioned against allowing unrelated content to seep in. Many educators do not see academic content as related to the pre-symbolic population. Within the development process, think about the unintended consequences. For example, once younger children are assigned the AELPA label, they might be tracked somewhere they were never intended to be.

A participant comment that there is a lot of good information in the examples in the Guide, but it's not a matter of just checking them off. She encouraged everyone to think about the body of evidence in terms of a validity argument. There are different levels of robustness for different pieces of evidence. There may need to be more combinations of evidence to make reasonable and sound arguments. Consider whether you are presenting a reasonable body of evidence to support your claims.

A great deal of expertise is available on students with significant cognitive disabilities, and states can use this expertise to move the work forward. It's important to focus on communication and language, to identify resources for AELPA, and to have people with different perspectives work together. Technical manuals developed in the alternate assessment consortium can be a starting point.

The World-Class Instructional Design and Assessment (WIDA) and Texas have alternate assessments; states can talk to them. The Alternate English Language Learner Assessment (ALTELLA) project aims to apply lessons learned from research on successful instructional practices, accommodations, and assessing ELs and students with cognitive disabilities to inform alternate English language proficiency assessments. They are also developing a template. The National Center on Educational Outcomes and English Language Proficiency Assessment for the 21st Century (ELPA21) and CCSSO are looking at the population's characteristics and what they need.

One panelist observed that states have resources they do not use well. The TACs would be very useful in helping with the development and design of assessments. She advised the states to ask TAC members for advice on validity. Networking is important. Meaningful collaboration can help provide input on a theory of action. Examples of resources are advocacy groups, agencies, and a wide variety of stakeholders who can assist in building capacity as states move forward. The Peer Review Guide includes several examples. This panelist also asked, "Why wait until the last minute?" Set up a repository and

arrange to gather evidence as you design the assessment, and be proactive, rather than having to go back years to look for evidence. If you gather evidence as you go, you will see where the gaps are.