Session 3F: Meeting the Requirements of Assessment Peer Review for IADA (Combined Session Between Focus Areas 1 and 3)

Panelists: *Kathy Banks, Carla Evans, Meagan Karvonen, Phoebe Winter*

Dr. Evans started the session by collecting questions via a QR code or URL, and then she shared several of them.

Will the conference convene a special panel of IADA peer reviewers with expertise in innovative assessment?

Dr. Evans replied that Dr. Peasley shared yesterday that ED is recruiting peer reviewers, as it typically does. IADA does not go through peer review until it becomes the state assessment. It has the same peer review requirements and will not require special training.

Does ED plan on providing reviewers any special guidance on flexibility to allow for innovation?

Dr. Banks replied that the CEs remain the same, but peer reviewers may receive some additional training. Make sure you clearly describe the fact that you captured the essence of the content standards in the test bank from which you selected items and there should not be a problem. There will most likely be webinars to help ensure that people can explain the process clearly to peers. Dr. Winter added that there are a lot of ways to meet all peer review CEs with an innovative assessment. If you have a good TOA and validity argument and you clearly explain why you are doing this and why it is different, then you have a better chance. Dr. Banks added that you want to streamline and avoid overwhelming the peers. ED is willing to provide individual assistance as you are getting ready for peer review.

Are there any components of peer review that are not already addressed in the progress reports?
Dr. Evans answered that she did not think so. In your final APR, feel free to ask whether you are providing the right amount of evidence for peer review.

[Session 3F Meeting the Requirements of Assessment Peer Review for IADA slides 10–15]
Dr. Karvonen explained that once in operational mode, IADA will require more procedures and evidence than you are used to thinking about, but she said the CEs can still be met. Key areas at play include alignment, reliability, comparability, and interpretation; she emphasized the importance of paying attention to the cross-references. Regarding alignment, Dr. Karvonen recommended focusing on coherent evidence as the broadest strategy for peer review. Your design should flow through into your blueprint and the way you think about test development and your operational pool. Be clear about your intended relationships, especially if they are inconsistent with the typical design approach. If you are doing performance-based assessment, it may not be automatically scored and probably has criteria and administration procedures that are all important parts of the chain of evidence in alignment. Think about ways to gather evidence of alignment as you go along so that you have it for peer review. For through-year

assessments, an alignment challenge is, “What is a window blueprint versus a master blueprint, and what are the expectations?”

Dr. Winter discussed reliability examples and challenges, stating that reliability, like alignment, always affects the system in design and scoring. Reliability in peer review comes up mostly in CE 4.1, but you also need to provide evidence of reliability in design and development and in scoring and reporting. When doing something innovative, support methods with research and explain why conventional methods wouldn’t work and give a rationale. In performance assessments, interrater training and agreement are important considerations and will most likely be more complex. Reliability scores, most likely across several dimensions, will need to be closely monitored. For through-year assessments, if end-of-year results are used to inform the next stages of instruction, the reliability of any single component may be more important than a summative score. The explanation is important; peer review asks how things work for the whole system, not how they work for the through-year part. Instructionally embedded assessments will be affected by how test scores are compiled and used to create summative tests. Watch out for the effects of retakes and various pathways on reliability.

Dr. Winter then discussed fairness and comparability. You want to ensure consistent, construct-related cognitive complexity across forms, full accessibility to all students regardless of forms, and equal familiarity from students with the texts, items, or tasks. Fairness of design is always crucial for comparability for all assessment methods. Consistent timing of performance assessments in relation to instruction should be considered to ensure fairness across classrooms for all types of assessments. Teachers and administrators must have a thorough understanding of how to administer components, including all accessibility provisions. A student’s score should not be affected by variations in the system.

Dr. Winter explained that interpretations of summative scores are part of peer review. For any test, frequently checking design and development decisions against the TOA or other statements can support appropriate score interpretations. Particularly for new reports, educators should be involved in the design of the reports rather than at late stages of development. Building in processes for revision is also of great importance. Other considerations include understanding how system characteristics (e.g., weighting) affect interpretations and clearly explaining how scores can be interpreted and why. Dr. Winter urged attendees to remember that design is critical and that it should match what you want the test to tell you.

[Session 3F Meeting the Requirements of Assessment Peer Review for IADA slides 16–25]
Drs. Evans and Karvonen provided brief overviews of state programs for attendees to discuss in small groups and identify which of the peer review elements would be challenging to meet. (Descriptions of the programs below are abbreviated.)

New Hampshire: Grade book data and a teacher judgment survey were used to give students an annual determination on a scale of 1 to 4; reports were sent to parents.

Issues identified by the group: One issue is comparability across schools. If every classroom is doing something different, how do we know whether there is comparability within schools, within ELA, and across systems? At least three tiers of comparability evidence need to be collected. Clear parameters are needed about who is included. Additionally, if students miss the common assessment, there could be missing data, creating a small sample size and preventing comparability analysis. Dr. Evans noted that there are clear rules in place about these issues. There is no test security. Dr. Evans noted that it also has to be taken into consideration that the schools are competency-based and that the districts have different score ranges for their classrooms. States must consider the auditing and training processes that need to be in place to ensure sameness. Classroom assessment maps need to be provided to ensure alignment to state content standards.

Louisiana: Students have the opportunity for instruction-informed assessment, a through-year model with innovative assessment tied to instruction. Cumulative assessments throughout the year produce a summative score. There is an eight-item linking set during each of the three testing windows. The innovation aims to improve the quality of alignment.

Issues identified by the group: Interpretation is a problem. For example, what do students know during the school year versus at the end of the school year? All three assessments are rolled up and not weighted. If the first score is low, it could hinder the student. Regarding creating a summative score, participants asked: How is the state going to use information from the various three timepoints for the through-year assessment as part of students' summative determinations of proficiency? Will the state need to provide information that students were not differentially affected if they took unit assessments? Will the state need to show that students with disabilities or English learners were not differentially affected when information from earlier in the year was incorporated into end-of-year summative scores?

The state needs to show evidence about how it is providing accommodations for students with disabilities and English learners as part of the design and administration because these are more like classroom-embedded assessments based on hot reads. Louisiana could face alignment issues, depending on the standards coverage of each of its through-year components.

North Carolina: The state proposed using the results from the fall and winter interim tests as "Stage 1"—basically an optional through-year assessment system, with the state test at the end of the year holding all the weight. Dr. Banks noted that peer training needs to be refreshed to broaden the peers' minds about different ways people are innovating.

Issues identified by the group: Equitability and comparability of data provided with different options could be a problem. If tests are shortened, it can reduce reliability, but it appears the issues have been worked out. Dr. Mbella replied that it is not an issue because the multistage adaptive is not shortened; each of the options is the same length. There are few peer review issues because the end-of-year test is basically the same as before. However, earlier information can inform a student's placement in the multistage adaptive assessment. The state just needs to show that the adaptive routing won't prevent a student from being accurately classified on the

summative assessment. The state also must show that it meets the requirements related to all the other typical peer review elements.

Georgia: The state ran two pilots, even though the law is clear that by the end of the demonstration period, there can be only one state assessment system. (1) Navvy is a standards-based assessment happening over the course of the school year, using diagnostic classification modeling (DCM) to define "mastery" on assessments tied to each standard in math and ELA. There are 20–30 standards per content area per grade. (2) MAP Growth is based on the Northwest Evaluation Association's Measures of Academic Progress (MAP). This is somewhat like the North Carolina model involving testing three times per year, potentially using the multistage design to replace the single summative test.

Issues identified by the group for MAP Growth: Regarding alignment, mapping items onto a coherent blueprint that would measure Georgia's standards in a meaningful way would have been a challenge. Accommodations need to be appropriate for the assessment design and range of students served. Missing data could make the assessment difficult to score.

Issues identified by the group for Navvy: The Georgia Department of Education would need an entire team on call to handle administration. Not all schools will allow the same number of attempts. What if students score lower on additional attempts? A key question is, How much is too much when it comes to missing data while trying to produce a summative assessment?

Participants did not discuss an overview of the Massachusetts program because of time limitations.

Questions and Comments

How do through-year models handle students changing and developing? They change throughout the year.

Dr. Evans replied that it depends on the claim you are trying to make (performance at the time of the assessment or by the end of the year). You need the evidence to support the claim your assessment is making. Most through-year models don't use information throughout the year in a student summative determination. A paper on the Center for Assessment's website (nciea.org/library/through-year-assessment-ten-key-considerations/) addresses these issues. Dr. Winter added that having rules and instructions on when a teacher can give assessments is critical, as is building in retesting and overlap, so that you can make the needed inferences. You will also most likely figure out how to weight each assessment.