Session 1E: Test Administration

Panelists: *Phoebe Winter, Nathan Dadey*

[Session 1E Test Administration slides 4–15]

Dr. Dadey remarked that the field is currently clarifying the boundaries of peer review, so it is understandable that people have concerns. It is the state of the field and is a work in progress. After providing an overview of the session, he identified the CEs covered: 2.3 (Test Administration), 2.4 (Monitoring Test Administration), 2.5 (Test Security), 5.1 (Students with Disabilities Inclusion), 5.2 (English Learners Inclusion), and 5.3 (Accommodations). These elements ensure that students have a fair opportunity to show what they know and are able to do. As a function of this, the inferences made about what students know and can do are defensible. Doing so involves ensuring that testing conditions are consistently standardized, secure, and monitored. Test irregularities must be minimized and addressed if they occurred. States must also ensure that all students can participate and show what they know and can do—including students with disabilities and English learners—through inclusion policy and accommodations. Dr. Dadey emphasized that standardized testing does not mean that every student gets the same test at the same time under the exact same conditions. Rather, variation in the tested content and conditions of measurement are purposefully connected to the inferences to be made about students. Therefore, test administration procedures differ, but an argument defends that difference.

CE 2.3 incorporates policies and procedures for standardized test administration. Standardization requires states to develop materials and procedures for (a) test distribution and administration, (b) documentation of irregularities and (c) requesting and receiving accommodations. These policies must be clear and clearly communicated. States must also train test coordinators and administrators. Those who use CAT must meet technology requirements, provide instructions for the use of technology in administration, and offer solutions to technology issues and ensure that clear contingency plans are communicated. To provide evidence for this CE, the state can submit test administration and other types of manuals (e.g., training and accommodations), training materials, and technology specifications. It is crucial to show that training actually took place by submitting attendance records. Dr. Dadey noted that as test administration becomes more complex, no amount of training is enough. He suggested that states move trainings to summer or the back-to-school period and provide materials. Early parent–teacher conferences can be a good time to support testing programs. Educators need to know about any changes in test administration as early as possible.

CE 2.4 (Monitoring Administration) covers requirements for the state to monitor the administration of its assessments to ensure that standardized test administration procedures are implemented with fidelity across districts and individual schools. Evidence can include monitoring procedures and records (e.g., relevant communications to individual schools and school districts), self-reporting, and help desk reports. In complex administrative conditions, states may rely on data analytics to obtain a rolling analysis of who has taken the assessment. For CE 2.5 (Security), the state must implement and document an appropriate set of policies and procedures to prevent test irregularities and ensure the integrity of test results. It needs to

show that security procedures are in place to prevent, detect, and investigate and remediate any test administration irregularities. Evidence for this CE might include forensic analysis and investigation procedures. Often, states do not have comprehensive policies in security or do not demonstrate that the policies and practices were implemented.

For CE 5.3 (Accommodations), states must make available appropriate accommodations and ensure that their assessments are accessible to students with disabilities and English learners. Evidence might include an accommodations manual, training materials, official communications, and audit trails. States should ensure that they demonstrate that students who need accommodations have received them. If there have been departures from parallel forms, states should explain why with supporting evidence.

[Session 1E Test Administration slides 16–23]
Consistency in test administration is key to attaining standardization when using multiple approaches. Dr. Warner explained that this consistent administration must be planned in the design and requires collaboration between the state and others. The state should provide peer reviewers with a detailed definition of "consistent administration"—including the allowable variations in the administration process. These variations may depart from previous practice. Challenges arise when the timing or content of administrations varies from one student to the next. For example, students could take multiple testlets after being instructed on a topic based on educator judgment. The process would be standardized through training even though administration is not. States will need to reconsider consistent administration when their assessments change (e.g., when they scale up current practices or develop new practices). Dr. Warner noted that states must always consider, monitor, and track the burden of test administration. He emphasized the importance of training and added that it may differ depending on the approach and subject area. States should consider this issue early in planning and discuss the trade-offs with stakeholders. This is an opportunity to discuss the needs of everyone involved in test administration and to make the case that a new system is beneficial.

Although best practices for administering the types of assessments currently seen under ESSA are somewhat established, research on fair administration policies for new types of assessment is ongoing. Traditionally, these policies and practices have focused on standardization in the sense that all students take the same test in the same way (with accommodations). But new approaches to assessment most likely require shifts in administration policies and practices that ensure that the intended interpretations are supported by the way the assessment is administered to individual students. When there are multiple test forms, the state must ensure that students take the right test at the appropriate time and that students are supported so the assessment yields a comparable interpretation of the learning standards.

[Session 1E Test Administration slides 24–32]
In the context of TYA, Ms. Rozunick explained that appropriate administration policies and practices must be well considered and planned, as accommodations must be incorporated at the three testing times. Policies and practices could include consistent provision of accommodations (with decision rules to make changes, if appropriate) and collection of classroom identifiers to facilitate reporting. They also include tracking missing data throughout the year (within and across schools) and thoughtful consideration about makeups (both during the year and at the end of the year). States might generate decision rules about what constitutes a valid case or score and significant help desk–style support throughout the year. Finally, states must consider the implications of TYA for exposure to questions during earlier test administrations, which might compromise later assessments. Regarding missing data, the administration policy should be designed to obtain information that is as complete as possible and include specifications about makeup testing with documentation. Additionally, states should outline the rules or adjustment for addressing missing data, including rules about the definitions of participation and nonparticipation. For example, TTAP offers three distinct testing opportunities, so Texas assumes that there will be missing data and is investigating possible comprehensive or summed scores. Operationally, the state currently offers only retest opportunities for high-stakes, end-of-course assessments for high school.

States should ensure that special rules for mobile students are in place, accounting for each possible type of mobile student in the rules—for example, students who were absent during the entire testing window or who took the test but received an invalidated result because of incompletion or void assessment. Ms. Rozunick suggested that states use a decision tree for what should be done in complex situations rather than formulate a policy on mobile students.

TYA also requires that states consider the effects of within-year exposure to content and whether it compromises later test administration—as might be the case with memorable items in a set of questions about a particular passage. As an example of how a state might address this issue, Ms. Rozunick explained that TTAP uses a multistage adaptive model. All testlets for a given year are created at one time to ensure that item coverage is similar across all three testing opportunities. Once an item is used for one test, it may not be reused for another that year. The state's requirement to release 100 percent of its items does not yet apply to this pilot program.

[Session 1E Test Administration slides 33–40]
Dr. Warner discussed appropriate administration policies and practices for performance assessment, which could include guidelines for how to integrate this kind of testing into ongoing instruction. They might also include rules for consistent distribution and collection of materials and support for determining appropriate student accommodations based on the specific attributes of the assessment and the student's needs. Such policies and practices could also include clear directions on which administration procedures must be followed exactly and where local flexibility is allowed without affecting the interpretations. He noted that New York's science performance assessments in lower grade levels are flexible (e.g., when it comes to the time of year the assessments are given) and support quality instruction. The administration policies are loose. The activities are aligned with learning content. Scoring rubrics are provided.

And the use of results is a local decision. Science performance assessments are not so flexible at the high school level because of course credentials. Integration of the performance assessment inside the content area is critical for consistent administration. The state ensures that the materials distributed are consistent and determines appropriate accommodations—which requires planning and collaboration with proctors for correct interpretation.

Matrix sampling approaches pose the fewest challenges, as they are similar to the current assessments administered under ESSA. Many of the traditional policies and practices are in place. States should consider developing a sampling plan to ensure that content coverage is appropriate and comprehensive across students and enables inferences at the subgroup level. They should also consider establishing limitations to testing windows to minimize instructional timing effects when results are aggregated. States may also want to develop audit procedures to examine the influence of differences in instruction across districts and other factors on the aggregated results. Audit procedure can be used for iterative improvement in test administration (e.g., reducing the burden on schools).

Dr. Warner commented that peer review is only concerned with assessment that produced summative annual determinations, which extends to test administration. States should check the polices and consider this from the beginning when designing their assessment systems. He stressed the importance of aligning assessment with purpose and use and defining consistent administration. For example, in a program in which multiple within-year assessments are given based on educator judgment, evidence for peer review might include a rationale of how consistent administration supports the validity argument. Other evidence might include clear boundaries, guardrails, and training on consistent administration and documentation that the process was implemented with fidelity. Multiple within-year assessments may increase the need for monitoring. Alternative approaches—such as sampling and technology—may be used to alleviate the burden of increased monitoring. States should demonstrate to peer reviewers that they have implemented monitoring and provide evidence (e.g., checklists) that links back to assessment design.

Questions and Comments

David Brauer of the Ohio Department of Education asked whether there is an acceptable or average percentage for remote monitoring for states with a significant number of LEAs. Ohio has about 700 school districts. Dr. Warner responded that he understands the challenge, as New York has almost 800 school districts and many independent schools. He suggested making a validity argument for a feasible sampling plan to capture the different types of school districts (e.g., rural and urban). There is no specific percentage, but the state could develop a multiyear plan to show that it will monitor as many districts as possible and that those districts are representative. States conduct various types of monitoring (e.g., accommodations), which can be done at the same time as assessment monitoring for efficiency. Dr. Banks of ED added that ED receives multiyear monitoring plans in which states explain that within five years, they will have monitored a target number of districts. Successful plans present a thorough description of the monitoring process and evidence that monitoring occurred (i.e., a sample of redacted

completed forms). When a district cannot be monitored, the state may submit communications that provide a through line of the process.

How can states that implement through-year testing ensure that the content on the fall test has been taught by that point in the school year?

Ms. Rozunick responded that states cannot dictate the content or timing of what school districts teach. Texas did consider this issue when designing TTAP and provides a blueprint.

What considerations (e.g., subject or research) go into different testing conditions and supporting inferences? How should states decide what can and cannot vary?

Dr. Dadey commented that the theory of action and values should be the foundation for assessment, followed by good arguments for points of departure and flexibilities. States should define and defend allowable variations within the inferences they want to make. Dr. Winter added that validity can be based on the views of content experts, or states may conduct testing to validate fairness and comparability claims.