Session 1A: Test Design and Development

Panelists: *Meagan Karvonen (karvonen@ku.edu), Nathan Dadey (ndadey@nciea.org), Brooke Nash (bnash@ku.edu), Christine Rozunick (christine.rozunick@tea.texas.gov)*

[Session 1A Test Design and Development slides 3–16]
Dr. Karvonen provided an overview of the session, noting that although subject matter experts are knowledgeable, they do not have all the answers and do not always agree with one another. Nevertheless, they have been considering various design options and can provide information for states to consider. She pointed participants to ED's 2018 guidance on peer review, located on the conference website. The session topic is primarily addressed in CEs 2.1 (Test Design and Development), 2.2 (Item Development), and 3.1 (Overall Validity).

Dr. Karvonen explained that CE 2.1 (Test Design and Development) aims to ensure that the state's test design and test development process are well suited for the content and are technically sound. To address this CE, the peer review submission needs to align assessments to (1) the depth and breadth of the state's academic content standards for the grade that is being assessed and (2) the depth and breadth of the state's grade-level academic content standards in terms of balance of content (i.e., knowledge, cognitive process, and cognitive complexity). CE 2.1 covers the assessment's statement of purposes and intended uses and blueprints that support test development (depth and breadth). It addresses whether the assessment is tailored to knowledge and skills in the standards, including appropriately complex applications. CE 2.1 includes the computerized adaptive testing (CAT) item pool and the selection procedures that support test design. Although not the focus of this session, CE 2.1 also includes English language proficiency (ELP) requirements.

For CE 2.2 (Item Development), peer review ensures that the state uses reasonable and technically sound procedures to develop and select items to assess student achievement. These must be based on the state's academic content standards (content as well as cognitive process, including higher-order thinking skills). CE 3.1 (Overall Validity) covers whether the state has documented adequate overall validity for its assessments in terms of nationally recognized professional and technical testing standards. Dr. Karvonen noted that there are four types of validity evidence and that evidence needs to carry forward from the previous peer review criteria. This CE covers whether assessments measure the knowledge and skills specified in the state's academic content standards. Specifically, it ensures adequate alignment between assessments and the academic content standards the assessments are designed to measure in terms of content (i.e., knowledge and process), balance of content, and cognitive complexity. It also ensures that assessments address the depth and breadth of the content standards. Dr. Karvonen noted that the session does not cover the alignment of assessments to content standards for AA-AAAS. The CEs addressed in this session may affect other CEs. CEs directly affected include 4.2 (Fairness and Accessibility), 4.6 (Multiple Versions), 4.7 (Technical Analysis, Ongoing Maintenance), and 5.3 (Accommodations). They may indirectly influence CEs 4.1 (Reliability), 4.4 (Scoring), and 6.4 (Reporting).

Dr. Karvonen commented that from the peer reviewer perspective, a challenge occurs when the state has met a requirement but does not present evidence for it (unintended gaps). For example, test procedures affect fairness but may not be presented or referred to in this CE. She noted that the examples included in this session would be based on variants of through-year designs and that other designs could be discussed during the Q&A.

In a through-year assessment (TYA) or instructionally embedded (IE) program, the assessment is administered in multiple sessions during a school year. TYAs are intended to support both the production of and the use of a summative determination and one or more additional aims. Possible TYA or IE designs include full domain (covers all the standards), modular (covers a subset of standards), and modular and flexible. Often, but not always, TYA/IE is associated with a design that has three administration windows, has an item response theory (IRT)–based vertical scale, and varies the use of the results from the first two windows in the summative score. States that use this design include Alaska, Nebraska, Maine, and Virginia. In the modular design, the state administers three assessments with an IRT-based scale, with each covering a partially unique subset of standards. This design features numerous assessments based on diagnostic classification models in which each assessment covers an individual or small subset of standards. States that use this design include North Carolina and Montana, as well as the Dynamic Learning Maps (DLM) Consortium. Dr. Karvonen provided the QR code for the Multiple Approaches Handout (Framing Table for Multiple Approaches to Assessment Design), which provided a summary of the assessment programs featured in Focus Area 1 sessions and slides related to Focus Area 1 from the plenary session.

[Session 1A Test Design and Development slides 17–21]
Dr. Nash focused her remarks on test design and development, covering issues that arise and general advice for addressing them. Common issues include: the state has too many and/or unclear purposes and uses as they relate to the academic content standards or the purposes and uses of the test design may not align well with the goals of the new or different assessment approach. States may adopt a new assessment approach in order to reduce testing time, but there is a trade-off with the ability to derive sub-scores for individual students. Additionally, the state may make a test design choice that does not align well with its purposes and uses.

To address these issues, states might use a theory of action (or another type of logic model) to identify the goals and long-term outcomes of the new or different assessment approach. States should articulate clear statements of intended interpretations and uses of assessment results. Dr. Nash emphasized the importance of obtaining agreement from stakeholders on the purposes and uses of assessments prior to designing tests. It is a good idea to focus on the connections between intended uses and each component of test design. This may include the purposes outside the purview of peer review that affect design and involve working iteratively from results-reporting designs (e.g., score reports and reporting dashboards). Dr. Nash suggested that if possible, states should conduct a small-scale pilot study to evaluate the test prior to full-scale development. The goal is to determine whether the design meets intended uses (e.g., guiding the next steps in instruction).

[Session 1A Test Design and Development slides 22–26]
Ms. Rozunick discussed common issues that arise with test design blueprints and ways to address them. States may not fully understand that the domain or content structure can lead to unnecessary blueprint requirements and may misunderstand how best to achieve depth and breadth of content standards. She emphasized that this can be difficult without appropriate planning and suggested that states develop a written plan and have others review it prior to finalizing the test design blueprint. States may not appropriately address the depth and breadth of content standards at each level of intended inference (e.g., within individual forms and across school districts, individual schools, and years). This requires them to consider the local pacing of testing and opportunity to learn (OTL) factors. Another issue that can occur is that release or reuse requirements for test items affect the blueprint. These requirements vary considerably; some states never release any items, whereas some states must release all items by law. Determining the number of administrations and test length is a common issue, and states must decide on the trade-offs (e.g., reduced or no sub-score reporting for within-year assessments) and key questions (e.g., How short is too short?)

To address these common issues, Ms. Rozunick suggested that states develop a clear description of the content structure and how it relates to the blueprint specifications. They should consider models and administration and the frequency of revisiting the plan (monitoring and adjustment). Early in the process, states should determine which parts of which assessments will be used in summative calculations—with a focus on what is essential for peer review. She added that blueprints can specify depth and breadth of content standards for the "total assessment" (i.e., for summative reporting) while still supporting other intended uses (e.g., reporting student skill mastery throughout the year).

When designing tests, states must also consider how they will be administered. Issues that often arise include that CAT's adaptive algorithms and procedures may need to be adjusted to accommodate changes in design. When the design includes multiple assessments, states will need to consider the distribution of test item pools. To address these issues, Ms. Rozunick suggested that states clearly define, test, and continually monitor test administration procedures and algorithms to ensure blueprint coverage is met at each intended level of inference. Early field monitoring will offer opportunities to adjust. It will be essential to work closely and early on with technology teams, vendors, and content developers on meeting the requirements for adaptive administration of multiple approaches to assessment.

[Session 1A Test Design and Development slides 27–30]
Dr. Dadey focused on test development, starting with procedures for item development. Common issues include the size and coverage of item pools needed to support multiple approaches to assessment. He emphasized that early design affects everything downstream, so states should consider gathering stakeholders to agree on item content prior to developing tests. Greater content integration is often needed and needs to link to the intended uses of the test. Comprehensively outlining test content and building checks into the process should help states avoid the common problems of misalignment of existing test item banks with state content standards or intended specifications of design. For example, an item written for a test

scored using IRT might not work well for a test scored using diagnostic classification modeling. States may need to adjust their item writer training and procedures to ensure that items meet the intended uses of the assessment.

To address the common issues related to item development, Dr. Dadey suggested testing the existing item bank against the design criteria early in the process. Additionally, it is crucial to consider the alignment methodology and criteria early in the process. States might consider using evidence-centered design task models to support item writing. When selecting items, states may find a mismatch between their procedures to meet test form requirements and the items available in the bank. Dr. Dadey emphasized that it is difficult when states ask content developers, coders, and others to adjust items, but this problem can be avoided through good planning and design. Defining item selection requirements first allows states to evaluate item bank and item development needs based on requirements and the item selection procedures that will be implemented.

[Session 1A Test Design and Development slides 31–36]
Ms. Rozunick highlighted the example of test design and development in the Texas Through-year Assessment Pilot (TTAP). Regarding TTAP's purpose and use, the state's goal was to develop a content-aligned, valid, and reliable assessment system that could replace existing assessments. TTAP is based in a well-considered theory of action. Texas has developed a progress monitoring system that provides timely data and information to support instruction. With TTAP, Texas aims for assessments that are minimally disruptive to instructional time, which is more complex than it appears. The state's primary goal is for TTAP to replace the State of Texas Assessments of Academic Readiness (STAAR), its current summative assessment. The TTAP team began each design consideration by reviewing current practice along with all possible options. The process also involved a review of legislative requirements prior to test design and blueprint development. A high-priority design consideration was assessment timing. Students have three opportunities for testing annually, with the first two distinct from the third. Therefore, students are not taking three summative assessments. The first two of these opportunities need to be shorter than the third.

Texas's approach to test design was to work in targeted grades across all levels and subjects (math, science, and social studies) to determine feasibility. Currently, the state is not moving to operational implementation. Texas is planning an RLA pilot of a couple of grades next year, which is particularly difficult under a multistage model. The intended score report information for TTAP is under development. Some basic reports have been created, but the state is reviewing them with stakeholders, who provide a great deal of input. The reports are static (and available online) but eventually will be dynamic. Texas is exploring cumulative scoring that would take into account the student proficiency demonstrated throughout the year, which would require research and policy considerations. A major issue has been the need for data literacy among educators who interact with the pilot. The program team explains the data and precursors and how they fit with the bigger picture. The pilot has received positive feedback.

[Session 1A Test Design and Development slides 37–49]

Dr. Nash discussed Dynamic Learning Maps (DLM), a year-end and IE assessment model. These assessments in English language arts and mathematics have been operationally administered in several states since 2014–15. The DLM alternate assessment system serves students with significant cognitive disabilities in grades 3–8 and high school. The results of DLM are intended to support interpretations about what students know and are able to do in each assessed content area. The results provide information that can be used to guide instructional decisions, as well as information appropriate for use with state accountability programs.

The DLM assessment model is based on learning maps that describe how students acquire knowledge and skills and provide a framework that supports inferences about student learning needs. DLM is based on evidence-centered design that includes a set of learning targets for instruction and assessment that is aligned with grade-level academic content standards and instructionally relevant assessments. DLM designs in accessibility and provides assessment results that are readily actionable and guide instruction. Although DLM is not a typical assessment program, it can be conducted and can meet all peer review requirements for use as an accountability assessment. Dr. Nash illustrated the DLM theory of action, noting that it aims to improve academic experiences and provide appropriate supports for educators. Test design and development decisions flow from the goal. She briefly reviewed DLM content structures and learning maps, in which essential elements (EEs) link to college and career readiness standards. For each EE, it is necessary to identify content standards with different levels of complexity in the blueprint.

In such an assessment system, students take testlets at instructionally relevant points in time across the school year. Testlets are based on nodes for one linkage level of one EE and contain three to nine items. A testlet begins with a non-scored engagement activity (e.g., context, a story, or information related to items in the testlet). Within DLM testlets, several item types are used (e.g., multiple-choice single-select and multiple-choice multiple-select).

Slide 43: Test Development Principles
The DLM system uses evidence-centered design procedures to develop test specifications and task templates for item creation that also incorporate universal design for learning (UDL) principles. The evidence-centered design approach is structured as a sequence of test development layers that include (a) domain analysis, (b) domain modeling, (c) conceptual assessment framework development, (d) assessment implementation, and (e) assessment delivery. Incorporating principles of UDL allows students to respond to items free of barriers while emphasizing accessibility and offering multiple ways to demonstrate skills.

Consistent with the theory of action, the assessment administration process reflects nonlinear and diverse ways that students learn and demonstrate their learning. Test administrators choose the content standards for assessment from the pool that meet a pre-specified set of criteria to achieve blueprint coverage. For each selected content standard, testlet administration procedures use multiple sources of information to assign testlets (linkage level), including student characteristics, prior performance, and educator judgment. Dr. Nash displayed a graphic describing relevant test design and development statements in the theory of action

and reviewed the relevant propositions (assumptions about the claim) and evidence for the appropriate combination of testlets.

The CGSA program offers Pathways for Instructionally Embedded Assessment (PIE). This competitive grant project began in 2022 and is led by the Missouri Department of Elementary and Secondary Education, in partnership with ATLAS. PIE is a four-year project aimed at designing, developing, and evaluating a prototype integrated assessment model for 5th grade general education students in mathematics. Grantees are pilot-testing new assessment ideas and concepts for potential use in the future. The features of PIE may include assessments based on learning maps (called learning pathways) and teacher selection of standards to create content groupings as the basis of instruction and assessment. The PIE pilot design includes full coverage of content standards in instructionally embedded assessment administration and end-of-year assessment administration.

[Session 1A Test Design and Development slides 50–57]
Dr. Karvonen discussed how states can respond to peer review requirements in test design and development, drawing upon lessons learned and suggestions on how to frame evidence. An overarching concern is that the current criteria do not work with innovative assessments; however, DLM counters this issue. She added that peers have knowledge and a traditional understanding of summative programs, so it is acceptable to explain innovative approaches. States should not assume that peers deeply understand the assessment design. States should include a succinct statement that "answers the question" related to the CE or provides the background needed to evaluate the evidence. Dr. Karvonen suggested that states leave "breadcrumbs" in the index responses to cross-reference CEs. It can be helpful to explain atypical evidence and show how the new aspect is similar to a familiar one. She emphasized the importance of coherence in the peer review submission and showed examples for particular CEs. States using innovative item types should explain the evidence of appropriateness to measure students' knowledge and skills and the depth and breadth of the standards (CE 2.1). For CE 2.2, states should provide adequate evidence that the test development procedures successfully produced those items. Cross-references help peer reviewers, and CEs 3.2, 4.2, and 5.3 offer opportunities.

States using an assessment system with multiple purposes and intended uses should provide evidence across the CEs that addresses how summative scores will be used. Submissions should clarify which parts of the system are subject to peer review when describing the evidence. States should delineate what is and is not subject to peer review (e.g., using different text formatting). Early in the process, conversations with the TAC can help states determine how to plan, evaluate, and synthesize validity evidence when they do not have a theory of action (CE 3.1). It is important to ensure consistent information across CEs 2.1, 2.2, and 3.1 (content evidence) and to crosswalk the information with CE 4.7 for areas with intended improvements.
Questions and Comments

David Brauer, English Language Proficiency Assessment Program Administrator at the Ohio Department of Education, asked about the typical rate of release of test items. In Ohio, they are

required by law to release 40 percent of test items annually. Ms. Rozunick responded that it ranges from 0 to 100 percent, depending on the state. Audra Ahumada, Deputy Associate Superintendent of Assessment at the Arizona Department of Education, wondered about the best practices for test item release. It may be better for a state to build up a large bank of test items before releasing a proportion. Zach Warner, Assistant Commissioner for the Office of Assessment at the New York State Department of Education, noted that high-release states are hindered from building up test item banks. It might be helpful for these states to release the document used to build the items instead.

Dr. Dadey focused on the purpose and use of assessments and asked panelists to provide examples of those that fall under peer review and those that do not. Dr. Karvonen commented that only information related to summative assessments is needed for peer review. Other purposes can include instructional guidance. However, she noted that evidence is often useful for multiple purposes. Ms. Rozunick added that some information on purpose and use can be used to ensure alignment. Dr. Dadey remarked that it is not always clear-cut whether information relates to one assessment purpose or another.