Session 1C: Fairness and Comparability

Panelists: *Phoebe Winter (phoebe.winter@outlook.com), Brooke Nash (bnash@ku.edu), Zach Warner (zachary.warner@nysed.gov), Meagan Karvonen (karvonen@ku.edu)*

[Session 1C Fairness and Comparability slides 4–18]
After providing an overview of the session, noting that another section of the conference would address IADA requirements for comparability, Dr. Winter reviewed fairness and comparability considerations for multiple approaches. Newly adopted approaches to assessment (e.g., TYA and matrix sampling) typically differ from traditional approaches in the inferences that they are designed to support. Newly adopted approaches may introduce between-student variation in the what, when, and how of assessment. The assessment approach is not new to the field, but its use in state assessment is novel. Dr. Winter stressed that comparability has to be at the level of inference that the state wants to make and that states should consider fairness and comparability throughout the process—from test design to score reporting. For example, states have to show how CAT is comparable to Braille and other forms of the test. She referred participants to the Multiple Approaches Handout.

Fairness and comparability are entwined. Although fairness and comparability affect all CEs, the session focused on CEs 4.2 (Fairness and Accessibility), 4.5 (Multiple Assessment Forms), and 4.6 (Multiple Versions of an Assessment). It was also applicable to CEs 2.1 (Test Design and Development), 3.1 (Validity), and 6.4 (Score Reporting). Some matters discussed in the session overlapped with those covered in others (e.g., validity). For all state academic and ELP assessments, assessments should be developed using the principles of universal design for learning (UDL), which is good for addressing peer review requirements. Dr. Winter suggested that states focus on relating UDL to specific areas of learning, ensuring that students have the necessary tools and familiarity for performance assessment, and reviewing content standards and their operationalization for all students.

For academic content assessments, the state must show that it has taken reasonable and appropriate steps to ensure that its assessments are accessible to all students and fair across student groups in their design, development, and analysis. If the state administers multiple forms of academic assessments within a content area and grade level, it must ensure that all forms adequately represent the academic content standards and yield consistent score interpretations. If the state administers any of its assessments in multiple versions within a subject area (e.g., online versus paper-based delivery or a Native language version of the academic content assessment), grade level, or school year, the state must show that it has (1) followed a design and development process to support comparable interpretations of results for students tested across the versions of the assessments and (2) documented adequate evidence of comparability of the meaning and interpretations of the assessment results.

Fairness is a necessary condition for score comparability. “Fairness” is defined as giving all students the opportunity to demonstrate the targeted knowledge, skills, and understandings. A fair assessment supports valid inferences that are comparable across the tested population. “Comparability” is defined as having scores that support inferences at the desired score level(s) and at the desired aggregation level(s). Both fairness and comparability involve a test with content that measures proficiency according to the same construct and whose results can be used for the same purposes regardless of test form or test conditions. For example, a state that tested knowledge of preassigned books must ensure that those texts are accessible in Braille.

In newly adopted assessment approaches, test design and development build in accessibility, feature clear and fair scoring rules and rubrics, and create multiple forms that are comparable. Dr. Winter emphasized that fairness and comparability must be designed proactively rather than addressed after review and that it is acceptable for a test not to produce sub-scores. Test scores that are comparable at the student level must cover the whole domain to provide a good estimate of what students know about that subject. The overarching goal is to ensure that standard inferences are made across examinees. Teachers need to know the appropriate time for an assessment, and students must know what is expected of them. Dr. Winter reviewed an example, showing considerations for assessment design and standardization that construct comparability. Generally, an approach that yields comparable results will support valid inferences for all students in the target population and have student forms designed to be as closely aligned to grade-level content standards as other forms. It will also provide results that are equally as reliable at the score level for which inferences are made and classify students into achievement levels based on the same degree of knowledge and skills.

[Session 1C Fairness and Comparability slides 19–26]
Dr. Nash discussed examples from states and consortia that illustrated how fairness and comparability should be considered in the IE approach and performance assessment. For IE assessment, fairness in design features year-round administrations that provide comparable contexts for students to demonstrate knowledge, skills, and understandings. The design minimizes construct-related variance and does not favor one group of students. It offers flexibility in administration and the ability to balance standardization with accessibility. Evidence of fairness in design for IE assessment may include high-quality and comprehensive test administration training and manuals and documentation of allowable (and nonallowable) practices. Test administration observations to evaluate fidelity and student cognitive labs are also acceptable evidence. For IE assessment, fairness in test development features large item pools to support an embedded approach and the use of evidence-centered design (ECD)–based task templates to ensure items are fair across student groups. Evidence of fairness in test development may include descriptions of ECD-based task template models and their development process.

In IE systems, forms represent content standards and feature short, embedded assessments based on multiple sources of information. Such assessments are dynamically generated and may only assess a few standards at a time, but all standards are eventually covered. States may show evidence that the forms represent content standards by providing peer reviewers with an analysis of the item pool to demonstrate breadth. To ensure that IE assessment forms yield consistent score interpretations, states can use ECD-based task templates to generate items that are written to precise cognitive specifications. Evidence may include an item data review demonstrating that items written to the same knowledge, skills, and understandings perform similarly.

[Session 1C Fairness and Comparability slides 27–29]
Dr. Warner explained that New York has a long-standing practice of including performance-based items in science examinations at all levels via the New York State Regents Examinations. More recently, the exams have incorporated curricular-embedded tasks. The science performance items allow students to demonstrate the specific knowledge and skills articulated in the learning standards through hands-on laboratory experiences as 15 percent of the total test score (combined with a written test). The state won a CSGA to explore the potential for its assessment strategy to be reimagined in a way that purposefully fosters high-quality instructional opportunities, provides authentic measures of deeper learning, and better prepares students for college and the workplace.

Fairness and comparability considerations in the design of performance assessments include focusing on the target(s) of measurement (rather than the task type) and ensuring the UDL is incorporated from the beginning. States also need to consider how they can balance flexibility in assessment with the need for consistent interpretations and access for all students. Another factor is the limitations of those who administer performance assessments. Dr. Warner emphasized the importance of a well-thought-out design in addressing these considerations. When fairness and comparability have been considered in design—including UDL-driven task development, a focus on knowledge and skills in PLDs, and decision documentation—these characteristics follow in the performance assessment system. The key is to ensure that students have what they need to show what they know and are able to do rather than give all students the same presentation and options.

The scoring of performance assessment should focus on targets of measurement and ensure that other aspects do not influence results. The scoring rubric must link to the values embedded in the assessment and the intended construct being measured. Assuming appropriate supports for students, score use should be tied to purpose. Reporting should clarify whether interpretations are affected by the flexibility in the assessment. Depending on the purpose and use of the assessment, comparability of scores may not be a reasonable goal. In that case, inferences would be standards-based and criterion-referenced.

[Session 1C Fairness and Comparability slides 30–32]
Dr. Winter focused on how states can prepare for peer review but noted that the line between what is in and out of the submission can be unclear. States with questions should check with ED. States can facilitate the understanding of peer reviewers by providing a good description of the assessment system and tying it to the intended purposes. Submissions should indicate why a component is included if it is not being presented for peer review (e.g., background information). The design section should refer to the purpose of the system and theory of action, tying these to evidence for relevant CEs. States should define how scores are used to meet ESSA requirements and tie this to the level of comparability to how scores are used for ESSA. Evidence of comparability may include the provision of appropriate supports and tools,

appropriate accommodations with backing (e.g., empirical studies and literature reviews), and audits. Dr. Winter added that when accommodations were first introduced, many people thought they were impractical. Given this history, it is exciting that the field is concerned about fairness and compatibility.

Dr. Karvonen presented an activity on the misperception of fairness and compatibility for participants.