Commentary on the U.S. Department of Education's Assessment Peer Review Process—

Increasing Innovation in Assessment

Susan Lyons

April 28, 2022

The purpose of this commentary is to identify and describe how the current federal peer review guidance places unnecessary restrictions on state assessment programs that limit the degree to which these programs can innovate in meaningful ways. While much of the same critique offered in this document can be applied more generally to the language of ESSA, the Critical Elements from the peer review guidance is used as an organizing structure.

This commentary defines innovation for state assessment programs broadly and recognizes that there are many possible valuable goals for innovation (e.g., reducing testing footprint, improving score estimates, developing tasks that more directly model good instructional tasks) and therefore does not explicitly prioritize one kind of innovation over another. Instead, this document points to the various ways that the current peer review guidelines could be limiting advancement in the field of state assessment, and in many cases, illustrative examples are provided.

The table beginning on the next page identifies potentially problematic language for innovation by Critical Element. The language address specifically by this comentary is highlighted and underlined.

Critical Element	Language from the Guidance	Commentary
2.1 (same comment applies to 3.1 as well)	"The State's test design and test development process is well-suited for the content, is technically sound, aligns the assessments to (1) **the depth and breadth of the State's academic content standards** for the grade that is being assessed; or (2) ***the depth and breadth of the State's ELP standards"***	The examples of evidence provided in the peer review guidance could acknowledge the reality that no assessment is reasonably able to demonstrate perfect alignment to the full depth and breadth of all the state-adopted content standards, but instead, each state must demonstrate how its content sampling plan (as articulated in the blueprint or analogous documentation) adequately supports accurate interpretations of student performance relative to the state expectations defined by the content standards. Some states may choose to prioritize breadth in content coverage at the sacrifice of measuring the full depth of the standards, while other states may instead focus on aligning to the full depth of a more limited, but important set of standards to better capture the cognitive complexity and higher-order thinking skills that the expectations require. Regardless of approach, each state should prepare a well-supported argument that justifies their sampling approach in relation to intended inferences about student performance relative to the full breadth and depth of the content standards.
2.3	The State implements policies and procedures for standardized test administration; specifically, the State: • Has established and communicates to educators clear, thorough and consistent standardized procedures for the administration of its assessments, including	Standardization in administration conditions is one way that assessments have traditionally sought to meet expectations of comparability in test score inferences. However, requiring strict standardization limits other avenues for gathering compelling evidence related to the comparability of score interpretations. Standardization is the operationalization of an outdated, positivistic notion that by standardizing the stimulus, administration conditions, and the response format, the student interactions with the test content are therefore also standardized and thus their behavioral item responses carry the same meaning about the student's level of proficiency relative to the content. This perspective fails to account for the latest consensus research on learning

Critical Element	Language from the Guidance	Commentary
	administration with accommodations;	that finds that student interactions with content are inherently dependent on their sociocultural contexts and the language-, experience-, identity-, and culturally-based structures of their knowledge schema (National Academies of Sciences, Engineering, and Medicine, 2018). Therefore, comparability cannot be meaningfully achieved by standardizing items and administration conditions. Instead, prominent measurement scholars are urging us to consider new pathways for achieving comparability by focusing less on standardization and more on evidence of alignment to the intended construct (Sireci, 2020; Mislevy, 2018; Moss, 1996; Pullin, 2008; Gee, 2008; Herman & Cook, 2019). With this perspective, comparability evidence would be aimed at demonstrating that each student has a comparable opportunity to meaningfully engage with the content to demonstrate their knowledge and skills. From a sociocultural perspective, a student is not presumed to have had the same opportunity to demonstrate their knowledge just because they are exposed to the same stimulus. Instead, the assessment environment must afford opportunities that attend to the learners' particular contexts (Gee, 2008). As Herman & Cook (2019) put it, "By better responding to student identity, culture, interests, and the interactive processes through which students develop capability, variations in the surface features of an assessment—such as holding students to the same criteria but permitting choice—may yield a better and fairer estimate of student capability" (p. 261).
2.5	The State has implemented and documented an appropriate set of policies and procedures to prevent test irregularities and ensure the integrity of test results through:	It will always be incumbent on the state to monitor, detect, investigate, and remediate any testing irregularities that could compromise the validity of its scores. However, specifying that test security must be obtained through a narrow conception of security of the materials and administration procedures limits the state's ability to meaningful innovate. For example, many

Critical Element	Language from the Guidance	Commentary
	• Prevention of any assessment irregularities, including maintaining the security of test materials (both during test development and at time of test administration), proper test preparation guidelines and administration procedures, incident-reporting procedures, consequences for confirmed violations of test security, and requirements for annual training at the district and school levels for all individuals involved in test administration;	certification programs have a long history of achieving test security by making the full pool of all possible items that could appear on the test available to potential test takers. This works well when the pool of items is so large that no examinee could reasonably memorize the correct answer to all the items without being highly proficient at the underlying skills. This runs against the theory that items must be strictly controlled to avoid cheating. It's reasonable to imagine that a state interested in developing a large pool of locally relevant performance-based assessments may curate a large and varied collection of performance tasks developed by educators across the state. A sample of these tasks may then be used on the state test. Test security could be maintained by not forecasting which performance tasks would be selected in advance of the assessment occasion. As always, the state would continue be responsible for collecting evidence that the scores of all students are comparable, with no potential unfair advantage or nefarious behavior.
3.3	The State has documented adequate validity evidence that the scoring and reporting structures of its assessments are consistent with the sub-domain structures of the State's (1) **academic content standards**;	The requirement that the scoring and reporting structures reflect the sub-domain structures of the State's content standards is unnecessary for supporting construct validity and favors a particular test design over a number of other viable alternatives. For example, validity based on internal structure can instead be demonstrated using a dimensionality analysis that reveals a strong, unidimensional factor representing the construct of the targeted grade and subject (e.g., Grade 4 math) or by using a sophisticated cognitive diagnostic modelling technique that validates student placement into achievement levels within the domain. The requirement to report and justify sub-scores associated with the sub-domains within the target grade and subject is limiting, and unnecessary for supporting the intended purposes and uses of Title 1 state

Critical Element	Language from the Guidance	Commentary
		assessment programs. To my knowledge, no state is utilizing sub-score reporting in their school accountability system for identifying schools in need of support. Nor are the sub-scores fine-grained enough for making meaningful decisions about student-level interventions. While there may be some limited potential uses of sub-score information for making programmatic decisions at the school or district level, the requirement that state assessment programs must report sub-scores is not necessary and interferes with the use of a number of innovative designs and measurement models that wouldn't readily lend themselves to this type of scoring and reporting.
4.2	**For academic content assessments**, the State 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.	The examples of evidence listed for this Critical Element should include evidence that the items have been designed with explicit attention to the diverse lived experiences and socio-cultural contexts of the students being served within the state. Leaders within the measurement community have been critical of traditional item development and bias and sensitivity procedures that result in items that favor the dominant cultural values and norms in our society, placing minoritized students at a disadvantage in relating to and responding to the stimuli (see Randall, 2021). State assessment programs should provide evidence that the assessment content is representative of the diversity of the cultural and linguistic groups within the state, with particular attention to representing the cultures and experiences of traditionally minoritized groups (e.g., students of color, immigrants, linguistic minorities, students with disabilities, etc.).
6.4	For **academic content assessments**, the State reports assessment results, including itemized score analyses, to	All of the current state assessments approved under Section 1111 fail to adequately meet this criterion when interpreted literally, and as a result of this language, test designs and individual student reports suffer (e.g., overly long,

Critical Element	Language from the Guidance	Commentary
	districts and schools so that parents, teachers, principals, and administrators can interpret the results and address the **specific academic needs of students**, and the State also provides interpretive guides to support appropriate uses of the assessment results.	overly confusing). Federal guidance should recognize that assessments designed to measure the full breadth and depth of the content standards for the purpose of school accountability are ill positioned to provide specific, diagnostic information about student-level academic needs and should not be required to do so. This requirement sends a counter-productive message to states and vendors seeking to minimize the footprint of state assessments while meeting the requirements, and also creates a false promise to the public about the value of these assessments to educators and students.

References

Gee, J. P. (2008). A sociocultural perspective on opportunity to learn. In P. A. Moss, D. C. Pullin, J. P. Gee, E. H. Haertel, & L. J. Young (Eds.), Assessment, equity, and opportunity to learn (pp. 76–108). Cambridge: Cambridge University Press.

Herman, J., & Cook, L. (2019). Fairness in classroom assessment. In S. M. Brookhart & J. H. McMillan (Eds.), *Classroom assessment and educational measurement* (pp. 243–264). Routledge.

Mislevy, R. J. (2018). Sociocognitive foundations of educational measurement. Routledge.

Moss, P. A. (1996). Enlarging the dialogue in educational measurement: Voices from interpretive research traditions. *Educational researcher, 25*(1), 20–29.

National Academies of Sciences, Engineering, and Medicine. (2018). *How people learn II: Learners, contexts, and cultures.* National Academies Press.

Pullin, D. C. (2008). Individualizing assessment and opportunity to learn. In P. A. Moss, D. C. Pullin, J. P. Gee, E. H. Haertel, & L. J. Young (Eds.), Assessment, equity, and opportunity to learn (pp. 76–108). Cambridge: Cambridge University Press.

Randall, J. (2021). Color-neutral" is not a thing: Redefining construct definition and representation through a justice-oriented critical antiracist lens. *Educational Measurement: Issues and Practice*, *40*(4), 82-90.

Sireci, S. G. (2020). Standardization and UNDERSTANDardization in Educational Assessment. *Educational Measurement: Issues and Practice, 39*(3), 100–105.