Enhancing Peer Review: Supporting Innovation in State Assessment Systems

October 2024

Lyons
ASSESSMENT
CONSULTING

FORESIGHT
LAW+POLICY

AUTHORS

Samuel D. Ihlenfeldt, Lyons Assessment Consulting
Sanford Student, University of Delaware
Susan Lyons, Lyons Assessment Consulting
Nathan Dadey, Center for Assessment
Ellen Forte, edCount, LLC
Phoebe Winter, Independent

CONTRIBUTORS

The recommendations provided in this report are based on the thoughtful contributions of the following list of assessment and measurement experts. These experts represent a range of perspectives and bring deep expertise in innovative assessment through their engagements with a broad set of diverse state assessment programs. Contributors are listed in alphabetical order.

Aneesha Badrinarayan, Learning Policy Institute
Laine Bradshaw, Pearson School
Maria D'Brot, FocalPoint Education
Charlie DePascale, Independent
Wei He, NWEA
Andrew Ho, Harvard Graduate School of Education
Meagan Karvonen, University of Kansas
Pohai Kukea Shultz, University of Hawai'i at Mānoa
Susan Lottridge, Cambium Learning
Rochelle Michel, Smarter Balanced
Stephen Sireci, University of Massachusetts

ACKNOWLEDGEMENTS

We are grateful for the ongoing thought partnership of Kirsten Carr and Matt Blomstedt of Foresight Law + Policy. Kirsten and Matt were key partners in conceptualizing and executing the work that led to the production of this report. Thank you also to the many practitioners who reviewed and provided feedback on this report including officials at ED, the members of the Technical Issues in Large Scale Assessment (TILSA) collaborative through the Council for Chief State School Officers, and policy partners at KnowledgeWorks, the Aurora Institute, the Center for Innovation in Education, and the Learning Policy Institute.

TABLE OF CONTENTS

INTRODUCTION

It has been almost a decade since the nation's landmark education bill was reauthorized as the *Every Student Succeeds Act*, which included a new provision authorizing states to pilot innovative assessment models "to meet the academic assessment and statewide accountability system requirements under Title I, Part A of the ESEA" (Office of Elementary and Secondary Education, Department of Education, 2016, p. 88940). Despite longstanding interest, few states have seized the opportunity to adopt innovative approaches. Experts consulted in the preparation of this report identified the peer review process as a commonly perceived barrier for states in pursuing novel approaches to state assessment. Recently, the United States Department of Education (ED) has made attempts to dispel states' perceptions of peer review as a barrier to innovation, evidenced by ED's 2023 state assessment conference in which invited participants illustrated how innovative models would work within the confines of peer review (see also Dadey, 2024). At that conference, department officials reminded states that the peer review process is flexible—that the evidence of technical quality should be adapted based on what is most appropriate for the design of the assessment. Innovative assessment programs that depart substantially from the status quo for statewide assessment programs (e.g., Dynamic Learning Maps Instructionally Embedded Alternate Assessment) have passed through the peer review processes, while others are in the piloting phase (e.g., Louisiana, Massachusetts, and North Carolina). Despite efforts from ED and the existence of operational innovative models, the perception of peer review as a barrier persists. States are hesitant to pour the necessary resources and political capital into designing, developing, and piloting a new assessment program that has the potential of failing to meet the rigorous standards for quality prescribed by the peer review guidance.

The purpose of this report is to encourage innovation in state assessment by providing recommendations on potential revisions to the peer review guidance (U.S. Department of Education [USED], 2018) that would address the unique evidentiary considerations for assessment models that depart from the status quo (e.g., in their design, administration, and/or scoring). The recommendations offered within this report do not redefine what is currently required for each critical element (i.e., the requirements within the left-hand boxes of the guidance); rather, they expand the current peer review guidance to include novel examples of evidence that meet those requirements (i.e., the right-hand side examples of evidence). We understand the important role peer review plays in ensuring states are developing and delivering high quality assessments as required by law. We do not endorse any weakening of the high standards that ED holds for states. Instead, we seek to encourage improvement and innovation by offering additional examples of evidence that are not currently listed in the guidance. Our intention is that by explicitly listing additional examples of evidence that depart from the standard assessment model[1], we can more directly communicate to states that there are a range of ways to meet the peer review critical elements. Many of the recommendations in this report, therefore, are not limited to innovative assessments; ultimately, we hope that these additional examples of evidence will allow states to improve their peer review submissions in general.

[1] By standard assessment model we are referring to the model for state assessment that has dominated the state assessment market for the past two decades. This model varies in its details across states but generally involves a single, end of year standardized assessment administration of primarily selected response items that are scored using an item response theory-based psychometric model.

PROCESS

Project Background

The Measurement Collective for Innovative State Assessments consists of groups of state, measurement, and policy leaders. The Collective aims to spur innovation in assessment and accountability within what is allowable under current law, but with an eye toward informing future policy. In winter and spring 2023, groups of state assessment leaders and measurement experts convened for three initial discussions about real and perceived barriers to innovation in state assessment. These initial conversations identified multiple opportunities to better support states interested in innovating their state assessment models. One of those opportunities is the language of the examples of evidence included in the peer review guidance (found at https://oese.ed.gov/files/2020/07/assessmentpeerreview.pdf and referenced throughout this report as the guidance). This report contains potential additions or revisions to the guidance's examples, in line with this initial suggestion from state assessment leaders and measurement experts.

Convening Experts

In early 2024, a panel of experts were identified by Lyons Assessment Consulting and Foresight Law + Policy as key partners and participants in a full-day working meeting on April 11, 2024 in Philadelphia, PA. There were a total of 18 attendees: 13 invited measurement experts, three facilitators from Lyons Assessment Consulting, and two participants from Foresight Law + Policy. Each of the invited experts was hand-selected for their expertise in measurement and psychometrics and their practical experience with the peer review process. The April convening of experts was grounded in a number of real illustrative examples of innovative assessment models that states are currently pursuing or interested in pursuing, and the tensions those models may present when submitting evidence to the peer review process.

Eliciting Feedback

Following the convening in April 2024, the authors of this report reviewed notes and artifacts to begin generating the set of recommendations presented below. All experts were invited to review this first draft on May 31, 2024. The draft report was updated in line with this feedback. Dr. Susan Lyons then presented the recommendations at the Technical Issues in Large Scale Assessment (TILSA) State Collaborative on Assessments and Student Standards of the Council of Chief State School Officers on June 27, 2024. Thirty-eight state leaders and partners submitted feedback on the recommendations, which has since been accounted for in the following draft recommendations. Additionally, we have consulted with officials at ED and policy partners at KnowledgeWorks, the Aurora Institute, the Center for Innovation in Education, and the Learning Policy Institute.

RECOMMENDATIONS AND SUGGESTED ADDITIONAL LANGUAGE FOR THE PEER REVIEW GUIDANCE

The following set of recommendations are provided in the following three categories:

I. Recommendations to Align the Peer Review Process with Its Intended Purpose
II. Recommendations to Surface Proven Opportunities to Innovate
III. Recommendations to Clarify Un(der)explored Opportunities to Innovate

The first category focuses on recommendations that would clarify what some aspects of peer review are intended to achieve (or prevent), and what the scope of the process is—this is particularly important for innovative assessments, but will also be helpful more broadly. Recommendations in the second category are focused on aspects of peer review where states have already had success innovating. The third category focuses on changes to make it clear to states that many innovations that have not yet been implemented in state testing would be compatible with the peer review process.

Each of the recommendations is designed to serve three mutually supportive purposes. First, to give confidence to states that the peer review process is flexible and can accommodate variations in assessment models; second, to improve the quality of state submissions by expanding on the types of evidence that would be appropriate to submit, particularly in cases where the existing examples of evidence in the peer review process may not directly apply; and third, to signal to peer reviewers that there are often many different legitimate ways to provide evidence in support of the critical elements, especially in the context of innovative models.

All of the recommendations follow the same structure in that we introduce the idea with a short description and rationale for the recommendation. That rationale is followed by a table that provides suggested language that ED could consider adopting or adapting within a revision of the current guidance. Throughout the tables in this report, the text in black reflects existing language within the peer review guidance, whereas suggested edits are presented in orange.

Recommendations to Align the Peer Review Process with Its Intended Purpose

Recommendation 1. Encourage states to clearly define which components of the assessment system are being submitted for review.

States that are administering assessments across multiple administration windows or test formats should clearly indicate which parts of the assessment system contribute to accountability metrics (e.g., achievement level classifications and, often, scale scores). Some assessment programs may include components that are not used to produce summative scores (e.g., formative items, testlet scores), and such components should be clearly named and excluded from peer review. This view is reflected within the current guidance, which notes that "… a State has the discretion to include in its assessment system components beyond the requirements of the ESEA, **which are not subject to assessment peer review**… A State also may include additional measures in its State assessment system, such as formative and interim assessments, which would not be subject to assessment peer review" (ED, 2018, p. 6, emphasis added).

We believe the current guidance could be bolstered by addressing the complexities of this issue directly within the examples of evidence. For example, the evidence could make it clear that when annual determinations are based on results of multiple assessment occasions, it is not the scores of each assessment that are subject to peer review—which are often provided to in an effort to support teaching and learning—but only the aggregate that is used to produce the annual determinations. In such cases, indicators of measurement quality (e.g., reliability, invariance, dimensionality) of each assessment occasion are not subject to peer review; only those of the summative score used to create the proficiency determination are relevant for review. However, there may also be situations where it is crucial for states to also provide evidence that each part of the whole meets focused technical criteria (e.g., in a two-stage adaptive test where the first assessment does not contribute to the summative determination, but is used to route into the second stage). The recommendation provided for Critical Element 1.3 applies broadly; specific recommendations related to reliability and validity are elaborated in subsequent recommendations.

Critical Element	Recommended Additional Example(s) of Evidence
1.3 – Required Assessments	Evidence to support this critical element for the State's assessment system includes … • If the assessment system includes components that do not contribute to the summative score, the State must provide clear documentation of which assessment components contribute to the summative annual determinations for all participating students (for example, an assessment system that is administered across separate occasions where a subset of occasions are relevant to the annual determinations, or an assessment system that blends summative and non-summative items where only selected items contribute information to the summative score). Note that assessment components that do not contribute to the summative annual determination are not subject to peer review.

Recommendation 2. Provide examples of how validity evidence can be used to support an argument about the appropriateness of the intended use of Title I assessments.

As currently structured, the guidance takes a "laundry list" approach to validity evidence, in which some of the types of validity evidence are named, but little consideration is given to the relative importance of each type of evidence or the ways that the evidence can be structured within an argument to make an overall claim about the appropriateness of the assessment for its intended uses. We recommend two updates to the guidance related to the submission of validity evidence.

First, the examples of evidence should include language describing the synthesis of validity evidence into an argument or set of claims. Second, expanding on Recommendation 1, the guidance should be clear that states should only submit validity evidence pertaining to the interpretations of scores that are relevant to Title I reporting of annual determinations for use in school accountability. Other uses of test scores that states may consider in developing their assessment systems (e.g., use in instruction) fall outside the purview of peer review. It should be clear to submitting states that they are not obliged to provide evidence relative to these interpretations.

Critical Element	Recommended Additional Example(s) of Evidence
3.1 - Overall Validity, Including Validity Based on Content	Evidence to document adequate overall validity evidence for the State's general academic and ELP assessments, AA-AAAS, AELPA includes documents such as: • A chapter on validity in the technical report for the State's assessments that states the purposes of the assessments and intended interpretations and uses, based on the summative annual determinations, and shows validity evidence for the summative uses of assessments that is generally consistent with expectations of current professional standards; ◆ This summary of intended interpretations and uses, and associated validity evidence, may be structured in a way that synthesizes the evidence and connects the evidence to the interpretations and/or uses that it supports, such as with a validity argument. ◆ The State should provide evidence, interpretations, and uses related to summative annual determinations for school accountability. Evidence for other uses of scores need not be included and shall not be evaluated by reviewers, though these uses may be noted in the rationale for the assessment system (Element 2.1).

Recommendation 3. Validity evidence related to internal structure should allow for the prioritization of the overall score.

The current examples of evidence suggest the dimensionality of the assessment should be consistent with the structure of the standards. While it makes sense for the blueprint to reflect the structure and emphases within the standards, the internal structure of the assessment as determined through empirical dimensionality analyses should prioritize the primary score interpretation. Dimensionality may be better considered as part of the alignment evaluation, rather than as a statistical issue, as some areas of the standards are appropriately underrepresented in the assessments given their less prioritized role in curriculum and instruction (e.g., they may be playing more of a support role for a larger grain size concept). Consequently, expecting enough items to produce a dimensionality structure that is consistent with the structure of the sub-domains could lead to substantial misalignment with overall curriculum and instruction targets. Over the past two decades, the most common assessment approach to scoring has been to estimate an overall score from a unidimensional, traditional (i.e., Rasch, two- or three-parameter logistic) Item Response Theory (IRT) model. Under this approach, we would expect to see a strong unidimensional structure to support the validity of the overall achievement score and the derived annual determinations. On the other hand, for programs leveraging categorical models, we would expect to see a different factor structure, which should be explained and supported in the submitted evidence.

Critical Element	Recommended Additional Example(s) of Evidence
3.3 - Validity Based on Internal Structure	Evidence to support this critical element for the State's general academic and ELP assessments includes: • Validity evidence based on the internal structure of the assessments that shows levels of validity generally consistent with expectations of current professional standards, such as: … ◆ Reports of analyses that show the dimensionality of the assessment is consistent with the dimensional structure implied by intended interpretations of the score(s) used for annual determinations under ESSA (e.g., evidence for unidimensionality if a single score per assessment is used for determinations; evidence of model appropriateness for multidimensional measurement approaches such as diagnostic classification models); ■ Dimensionality evidence pertaining to sub-scores should be reported only as applicable, with the strength of the evidence commensurate with the extent to which sub-scores are used for proficiency determinations. For the State's AA-AAAS and AELPA, evidence to support this critical element includes: • Validity evidence that shows levels of validity generally considered adequate by professional judgment regarding such assessments, such as: ◆ Reports of analyses that show the dimensionality of the assessment is consistent with the dimensional structure implied by intended interpretations of the score(s) used for annual determinations under ESSA (e.g., evidence for unidimensionality if a single score per assessment is used for determinations; evidence of model appropriateness for multidimensional measurement approaches such as DCMs); ■ As applicable, dimensionality evidence pertaining to sub-scores should be reported, with the strength of the evidence commensurate with the extent to which sub-scores are used for proficiency determinations.

Recommendations to Surface Proven Opportunities to Innovate

Recommendation 4. Acknowledge innovations in how summative scores are reported.

Annual determinations can take a variety of forms that extend beyond the common approach of achievement levels based on the categorization of scale scores, and there are many legitimate reasons why states may opt to submit just achievement levels or other non-numeric outcomes in lieu of traditional scale scores. As an example, the Dynamic Learning Maps assessment program employs diagnostic classification models (DCM) for scoring which produce mastery profiles that do not correspond to raw or scale scores. For this kind of approach, the evidence to support Critical Element 4.4 should explain in detail what each score entails and how it is created. As another example, an innovative assessment system may report a summative score that is constructed from multiple tasks administered throughout the year that involves transforming the multiple task scores into a summative scale score. In line with Recommendation 2, some assessment models may include aspects that are unscored within the summative assessment; in these instances, only the relevant, summative score is subject to review. The peer review guidance for Critical Element 4.4 could better support innovative programs by explicitly stating that these and other approaches to creating annual determinations meet federal requirements.

Critical Element	Recommended Additional Example(s) of Evidence
4.4 - Scoring	Evidence to support this critical element for the State's general academic and ELP assessments, AA-AAAS, and AELPA includes: • A chapter on scoring in a technical report for the assessments or other documentation that describes scoring procedures, including: ◆ An operational definition of what constitutes a score for summative reporting purposes (e.g., achievement levels, scale scores, composite scores, latent classes). ◆ Procedures for constructing scales used for reporting scores and the rationale for these procedures or if non-numeric scores are used, rationale and procedures used for deriving these scores.

Recommendation 5. Refer to a "set of performances" rather than a "performance continuum."
A central question of the peer review process is: Can all students taking the assessment, even very high- or low-performing students, engage with the assessment and receive a score with an acceptable level of measurement precision that reflects their proficiency in the target domain? While content representation is covered in other critical elements, Critical Element 4.3 is squarely concerned with measurement precision and cognitive complexity. The language in Critical Element 4.3 and the associated examples of evidence employ terminology that assumes a continuous scale score that may not reflect the scoring practices in programs that employ other models. For example, diagnostic classification models (DCM) provide profiles of mastery classifications. It would be inaccurate to portray these classifications as defined based on a single continuum. The current language in the guidance appears to require that precision be defined in terms of a continuum—which could be interpreted as discouraging the use of these kinds of models. Consequently, states may feel discouraged from submitting with this type of approach as it is not currently reflected in the examples of evidence. The suggested small language change provided in the table below is intended to be more inclusive of innovative assessment models with nontraditional approaches to scoring and reporting student performance.

Critical Element	Recommended Additional Example(s) of Evidence
4.3 - Full Performance Continuum	For the State's general academic and ELP assessments, evidence to support this critical element includes: … • Description of the distribution of cognitive (for academic assessments) or linguistic (for ELP assessments) complexity and item difficulty indices that demonstrate the items included in each assessment adequately cover the full set of performances ~~continuum specified in~~ described in the State's (1) challenging academic content standards; or (2) ELP standards; For the State's AA-AAAS and AELPA, evidence to support this critical element includes: … • For students at the lowest end of the performance range ~~continuum~~ (e.g., pre-symbolic language users or students with no consistent communicative competencies), evidence that the assessment system provides appropriate performance information;

Recommendation 6. Include considerations for all assessment models within the examples of reliability evidence.

As previously noted in Recommendation 1, if a score is not used for annual determination purposes, it is not within the purview of peer review, and this distinction must be made unambiguous to both reviewers and states. The guidance should make it clear that the overall score used for making annual determinations is the priority for demonstrating evidence of reliability, or—more generally—measurement precision, and that scores used for lower-stakes purposes—such as domain subscores or individual testlet scores—are not held to the same reliability standard.

Additionally, measurement precision evidence for approaches to less commonly used assessment scoring approaches such as diagnostic classification models (categorical IRT) does not closely resemble the type of reliability evidence typically produced for more common continuous scoring models. States should be encouraged to tailor their evidence to their model in a way that makes it clear to reviewers that best practices for the given model are being followed. To support states interested in departing from the most common approaches, the peer review guidance can include additional examples of evidence to signal that all types of scoring models will be fairly reviewed based on criteria appropriate for the given model.

Critical Element	Recommended Additional Example(s) of Evidence
4.1 - Reliability	Collectively, evidence for the State's general academic assessments, the general ELP assessments, the AA-AAAS and AELPA must document adequate reliability evidence generally consistent with nationally recognized professional and technical testing standards. ***For ELP assessments***, such evidence should also be provided for any domain or component sub-tests, if applicable. The strength of reliability evidence should be commensurate with the extent to which a given score is used for accountability purposes. Evidence to support this critical element for the State's academic content and ELP assessments includes documentation such as: … • For measurement models where common approaches to reliability such as Cronbach's *a* are inappropriate, a detailed explanation of how reliability is conceptualized for the given measurement model and what would be considered acceptably high reliability for a score produced by such a model. Where appropriate, simulation studies and/or citations to published work on the model in question may be included.

Recommendation 7. Note that innovative assessment programs may have a different set of relationships with external variables.

In this Critical Element 3.4, the peer review process should focus on evidence that includes an interpretation of whether and why scores from a given assessment should or should not correlate with common external variables such as other academic tests, consonant with professional standards for evidence from relationships with other variables. It is incumbent upon the submitting state to explain how these relationships support the validity of interpretations and uses of the scores and why, for example, discrepancies with another assessment may be an expected outcome for legitimate reasons. Experts noted, that it may be reasonable to expect that the scores resulting from innovative assessment programs will differ from the traditional testing systems they are designed to improve upon. Whether an assessment is innovative or not, it is also important to note that academic tests are just one of many types of "external variables" to which scores might be expected to relate. States must therefore plan carefully which variables are included in analyses, and should include a rationale for why it matters that the given assessment produces scores that reveal high, moderate, or low correlations with the selected variables.

Critical Element	Recommended Additional Example(s) of Evidence
3.4 - Validity Based on Relations to Other Variables	Evidence to support this critical element for the State's general academic content and ELP assessments includes validity evidence that shows the State's assessment scores are related as expected with criterion and other variables for all student groups, such as: … • Analyses and explanations for what patterns of correlations are reasonable and expected based on the program's design, including an explanation of why the given external variables were included in those analyses. For the State's AA-AAAS and the AELPA, evidence to support this critical element includes: • Validity evidence that shows levels of validity generally considered adequate by professional judgment regarding such assessments, such as: … • Analyses and explanations for what patterns of correlations are reasonable and expected based on the program's design, including an explanation of why the given external variables were included in those analyses.

Recommendation 8. Include considerations for AI-enabled scoring.

The infrastructure to support multilayer neural networks or other complex machine learning models (colloquially known as AI) has progressed considerably in recent years and use of these models is becoming more widespread. However, without explicit reference in the guidance, states may face challenges in defending their implementation due to their size, complexity, and lack of transparency. The suggested additions to the guidance enumerate best practices for leveraging AI scoring models and give states a strong footing to justify their use. The recommended language is intended to support states in providing the evidence that advanced automated scoring models can accurately and consistently rate student responses. Given the ever-changing nature of this novel set of approaches, we have attempted to construct a set of examples that do not reference any one specific family of models (e.g., LLMs).

Critical Element	Recommended Additional Example(s) of Evidence
4.4 - Scoring	Evidence to support this critical element for the State's general academic and ELP assessments, AA-AAAS, and AELPA includes: … • If State uses advanced automated scoring (i.e., scoring using a machine learning model): ◆ Description of model development, including rationale for the model selected, training data, and procedures for ensuring model is fair/unbiased; ◆ Procedures for monitoring ongoing model performance, (e.g., random verification samples, benchmarks); ◆ Evidence of ongoing changes to the model, including a description of a schedule of changes to the model, the nature of the changes, and ongoing validation efforts based on the updated model; ◆ Evidence of a contingency plan should the model fail to produce reliable scores in ongoing maintenance including backup scoring systems and procedures for retraining the model; ◆ Evidence of the protection of examinee data and item content; ◆ Evidence of consistent and reproducible scoring by the automated scoring. • If the state uses a hybrid approach that combines advanced automated scoring with human scoring: ◆ Procedures for routing responses to human scorers (e.g., proportion of responses scored by humans, adjudication methods) and the rationale for those procedures; ◆ Procedures for training/monitoring human performance (e.g., criteria for rater selection, documentation of training, benchmarks for assessing drift); ◆ Evidence that advanced automated scoring produces scores that are comparable to those produced by human scorers, such as rater agreement rates for human- and machine-scored samples of responses (e.g., by student characteristics such as varying academic achievement levels or ELP levels and student groups), systematic audits and rescores; • For machine or advanced automated scoring of constructed-response items or other novel item types: ◆ Evidence that the scoring algorithm and procedures are appropriate, such as descriptions of development and calibration, validation procedures, monitoring, and quality control procedures; ◆ Evidence that machine or advanced automated scoring produces scores that are comparable to those produced by human scorers, such as rater agreement rates for human- and machine-scored samples of responses (e.g., by student characteristics such as varying academic achievement levels or ELP levels and student groups), systematic audits and rescores;

Recommendation 9. Highlight the flexibility in how states choose to establish evidence of alignment. The experts we consulted for this report noted a widespread, though not ubiquitous, understanding among states that peer reviewers expect alignment studies to use the Webb alignment methodology, including its reliance on Depth of Knowledge (DOK) as the means for evaluating alignment of cognitive complexity. This appears to stem in part from the history of alignment and in part from language in the guidance emphasizing the importance of the "depth and breadth" of a state's content standards in its alignment study. While depth and breadth are not Webb-specific terms, but rather informal terms used to describe aspects of validity evidence based on test content (AERA, APA, & NCME, 2014), there is no getting around the fact that alignment is commonly viewed as a barrier to innovation because submitters may not know that there are multiple ways to evaluate alignment, and there is no requirement that states follow Webb's methodology.

It is therefore important for the guidance to make it clear that alignment evidence should be based on the design of the system and meet criteria appropriate for the design, especially in the case of innovative assessment systems. As alternatives to Webb's method, we note techniques produced by edCount (see Forte, 2017) and conceptualizations offered by the Center for Assessment (see Gong & Patelis, 2016); Webb's methods can also be modified in numerous ways. Forte's (2017) methodology centers achievement standards as a key lever in establishing evidence that assessments reflect the full range of achievement expectations; Forte also recommends using a complexity framework that addresses complexity of the item stimuli (e.g., Achieve; 2019), the processes necessary to generate a correct answer, and the processes necessary to record a response (e.g., selected-response, extended written response). Gong & Patelis (2016) suggests evaluating alignment using CCSSO's criteria for high-quality assessments. The suggested language in the table below helps clarify to states that they may submit evidence of alignment that reflects a range of research-based methodologies including, but not limited to, Webb's method and the other methods cited in this report.

Critical Element	Recommended Additional Example(s) of Evidence
3.1 - Overall Validity, Including Validity Based on Content	Evidence to document adequate validity based on content for the State’s general assessments includes: … • Evidence of alignment, including: ◆ Report of results of an independent alignment study that is technically sound (i.e., use of a research-based and appropriate method and process, appropriate units of analysis, appropriate grain size of analysis, use of clear and appropriate criteria; demonstration of adequate agreement for any evaluative judgements conducted by panels of reviewers) and documents adequate alignment, specifically that: ■ Evidence is based on the design of the assessment system and meets criteria appropriate for the design; ■ Each assessment meets ~~is aligned to its~~ test blueprint specifications, and each blueprint, as applicable based upon the methods underlying the alignment study, ~~and each blueprint addresses: (1) **depth and breadth of the State’s academic content standards; or (2)** ***the depth and breadth of the State’s ELP standards*** and~~ reflects a sufficient relationship with the clearly defined assessment targets to support intended score interpretations relative to the design of the assessment system. For the State’s AA-AAAS and AELPA, evidence to document adequate validity based on content includes: … • Evidence of alignment, such as: ◆ Report of results of an independent alignment study that is technically sound, with evidence appropriate to the design of the assessment system, and that document adequate linkage between each of the State’s assessments and the: (1) academic content the assessments are designed to measure; or (2) English language acquisition skills the assessments are designed to measure;

Recommendations to Clarify Un(der)explored Opportunities to Innovate

Recommendation 10. Include a test design rationale for innovative programs.

The review of evidence in relation to innovative assessment programs may benefit from the inclusion of an assessment design rationale, particularly as it relates to Critical Element 2.1: Test Design and Development. Often, states are seeking to innovate in their assessment model for reasons that are highly relevant to overall systemic change (e.g., provide more timely information to stakeholders, create coherence between instruction and assessment, support student learning through actionable feedback); these reasons for innovating lead to specific design decisions which inherently come with benefits and trade-offs, as with any design process. The inclusion of a design rationale would support states in better justifying their design decisions in relation to the overall program goals and constraints. Additional supporting evidence to submit in relation to key design decisions may include a theory of action, feedback from stakeholder engagement efforts, discussion notes and recommendations from technical advisory committees, and citations to relevant research literature.

Critical Element	Recommended Additional Example(s) of Evidence
2.1 – Test Design and Development	Evidence to support this critical element for all of the State's assessments includes: For the State's general **academic** content and ***ELP*** assessments: • Documentation of the overall structure of the assessment system, including terminology relevant to interpreting any of the evidence listed below; … • Documentation of the rationale for choices made during the design and development of the assessment, especially where such choices represent innovations or departures from common practice. This may include, for example, a theory of action, feedback on design decisions from stakeholder and technical expert input, and/or citations to research illustrating how the design produces more valid inferences or better supports student learning. For the State's AA-AAAS and AELPA: • Documentation of the overall structure of the assessment system, including terminology relevant to interpreting any of the evidence listed below; … • Description of the structure of the assessment, for example, in terms of the number of items, item types, the proportion of item types, response formats, types of scoring procedures, and applicable time limits. For an assessment that is partially administered through portfolios or includes extended performance tasks, the description should include the purpose and design of the portfolio or performance tasks, exemplars, artifacts, and scoring rubrics; ◆ Rationale for choices made during the design and development of the assessment, especially where such choices represent innovations or departures from common practice.

Recommendation 11. Attend to purposeful variation in test administration, monitoring, and security. Peer review requires that each state define and implement "clear, thorough and consistent standardized procedures for administration" (USED, 2018 p. 40). Standardized procedures for administration do not require that every student receives the exact same test under the exact same conditions of measurement. ESSA and its IASA and NCLB predecessors allow for variations in test administration for some students, recognizing that administration conditions can influence the ability of a student to demonstrate what they know and can do (see e.g., Buzick et al., 2023). In these cases, states need to define allowable variations in tested content and conditions of measurement, and ensure that they are purposefully connected to the inferences to be made about students.

Many innovative assessment designs seek to further personalize the assessment experience beyond providing variations for identified student needs (e.g., language, disability status) or administration modes (e.g., device). These variations from a single assessment model may seek to test students on (1) the same content at multiple times during the year, (2) the same content at different times during the year, or (3) somewhat different content at different times. For each of these designs, a state will need to first articulate what variations in content and timing are allowable, how these allowable variations support the intended inference, and how a consistent administration procedure has been developed and implemented to support these allowable variations.

Consider the previous example from Recommendation 3, in which students take many short testlets distributed throughout the year. Further suppose that each testlet aligns to a small part of the content domain and that there are different patterns of testlet administration. In this example, we now have groups of students being assessed on parts of the content domain at different times. For this program, the state should develop and articulate an administration plan to best support the intended inferences for the test. This may involve, for example:

- The state has defined the process for identifying when students have had sufficient instruction to be assessed, and has documented that process and communicated it to educators;
- Educators have received training on this identification process;
- Educators implement this process with fidelity; and
- The state monitors the administration for patterns that suggest educators are not implementing the process with fidelity or that patterns inappropriately vary across student groups.

The following suggested language provides additional examples of evidence to be considered for Critical Elements 2.3, 2.4 and 2.5, related to test administration, test administration monitoring, and test security, respectively.

Critical Element	Recommended Additional Example(s) of Evidence
2.3 - Test Administration	Evidence to support this critical element for all of the State's assessments includes: … • Regarding test administration: … ◆ For assessment systems made up of multiple within-year components with possible variations in the order and number of the administered components, documentation that the state has established and communicated clear expectations on the timing and ordering of all possible variations, procedures for assigning/modifying accommodations across administrations, and policies regarding students who do not participate in all components of the test.

2.4 - Monitoring Test Administration	Evidence to support this critical element for all of the State's assessments includes: … • Brief description of the State's approach to monitoring test administration (e.g., monitoring conducted by State staff, through regional centers, by districts with support from the State, or another approach), including evidence that monitoring plans match the structure of administration; … • If an assessment system has variations in the order and number of administered within-year components, evidence that the state has a process in place to detect unintended patterns of non-participation or test administration.
2.5 - Test Security	Collectively, evidence to support this critical element for all of the State's assessments must demonstrate that the State has implemented and documented an ~~appropriate~~ approach to test security appropriate to the design of the test. Evidence to support this critical element for the State's assessment system includes: … • Enumeration of all possible irregularities identified by the state relevant to the design of their assessment system, along with a description of how each is addressed in practice.

Recommendation 12. Value additional priorities in item development procedures.
States engaging in innovative assessment may provide additional evidence of quality related to item development including processes related to co-design with stakeholders, features of cultural relevance, or particular attention paid to instruction and curriculum in the item development process. These enhancements may strengthen overall system alignment and contribute to the validity of score interpretations. Peer review can signal the value of these innovations by including related examples of evidence in the guidance.

Additionally, item development need not occur at the standards level, but instead, items or item sets can be developed to measure learning outcomes of a different grain size, which ultimately can be mapped back to standards. For example, items could be developed to directly align with sets of well-developed claims, measurement targets, or aspects of a curriculum framework, rather than individual standards. High quality instructional materials and pedagogy may focus on concepts that combine standards, rather than individual standards; therefore, assessment practices that target these larger grain size concepts may be appropriate when these sorts of instructional shifts are desired.

Critical Element	Recommended Additional Example(s) of Evidence
2.1 – Test Design and Development	For the State's general academic content and ELP assessments: … • For assessments that incorporate additional design priorities, documentation of the approaches the State uses to ensure the test design and item types address the intended features of the assessment (e.g., cultural relevance, instructional relevance). For the State's AA-AAAS and AELPA: … • For innovative assessment designs, documentation of the approaches the State uses to ensure the test design and item types address the intended features of the assessment (e.g., cultural relevance, instructional relevance).
2.2 – Item Development	For the State's general academic content and ELP assessments, evidence, such as a section in the technical report for the assessments, that shows: … • For assessments that incorporate additional design priorities, evidence that item development processes drew upon diverse stakeholder input related to cultural relevance, curricular relevance, student experience, and other aspects of assessment quality in addition to alignment to academic content standards. • If the assessment targets are different from the state-adopted content standards (e.g., different grain size), thorough description of the processes and logic argument that links the set of items on a test event with the content standards.

3.1 – Overall Validity, Including Validity Based on Content	Evidence to document adequate validity based on content for the State's general assessments includes: • Validity evidence based on the assessment content that shows levels of validity generally consistent with expectations of current professional standards, such as: … ◆ As applicable, evidence that item content reflects the input of diverse stakeholders to support the relevance of the item content to culture, curriculum, and students' experiences in support of the test's theory of action and purpose. For the State's AA-AAAS and AELPA, evidence to document adequate validity based on content includes: … • Validity evidence that shows levels of validity generally considered adequate by professional judgment regarding such assessments, such as: … ◆ As applicable, evidence that item content reflects the input of diverse stakeholders to support the relevance of the item content to culture, curriculum, and students' experiences in support of the test's theory of action and purpose.
3.2 – Validity Based on Cognitive Processes/ Linguistic Processes	Evidence to support this critical element for the State's general academic content and ELP assessments includes: • Validity evidence based on: (1) for academic assessments, cognitive processes; or (2) for ELP assessments, linguistic processes; that show levels of validity generally consistent with expectations of current professional standards, such as: ◆ Results of cognitive labs exploring student performance on items that show: (1) for academic assessments, the items require complex demonstrations or applications of knowledge and skills as described by the standards, without interference from construct-irrelevant variance such as that caused by cultural irrelevance; or (2) for ELP assessments, the items require targeted demonstrations or applications of linguistic knowledge and skills; For the State's AA-AAAS and AELPA, evidence to support this critical element includes: • Validity evidence that shows levels of validity generally considered adequate by professional judgment regarding such assessments, such as: ◆ Results of cognitive labs exploring student performance on items that show the items require demonstrations or applications of knowledge and skills as described by the standards, without interference from construct-irrelevant variance such as that caused by cultural irrelevance;

Recommendation 13. For tests with multiple forms, comparability should be established at the level of the annual determinations.

Many states innovating in their assessment design are doing so to improve coherence with the local context. For example, an assessment may have multiple forms, each tailored to a particular curricular product, instructional sequence, or language of instruction. In the case of computer adaptive assessment, there are an almost infinite number of possible "forms" depending on a student's level of achievement and response patterns. Given these variabilities in assessment forms, the peer reviewers should be considering the comparability of the intended inference. In the case of Title 1 state assessments, the intended inference of consequence is the summative annual achievement-level determination. Strict scale score comparability, in the psychometric sense, may not be required given the intended use of statewide assessments for school accountability. In gathering evidence related to the consistency of the annual determinations across versions and forms, states should be considering the match of students to version (e.g., aligned with the instructional scope and sequence), the quality of each testing event (e.g., % of testing events that meet the requirements of the blueprint), and the comparability of the academic achievement standards.

Critical Element	Recommended Additional Example(s) of Evidence
4.5 - Multiple Assessment Forms	Evidence to support this critical element for the State's assessment system includes: … • Documentation to support the comparability of achievement level determinations: ◆ When applicable, documentation of the range of possible forms, including how the State ensures that the form(s) seen by each student meet(s) the content blueprint requirements. ◆ When applicable, procedures for interpreting and comparing scores when administration conditions differ meaningfully (e.g., number of administrations, administration windows, curricula), producing atypical psychometric properties (e.g., non-invariance of IRT item parameters), and rationale for those procedures.
4.6 - Multiple Versions of an Assessment	Evidence to support this critical element for all of the State's assessments includes: • When applicable, procedures for assigning students to a set of individually appropriate tasks (e.g., number of tasks, language of instruction, curriculum relevance) and a justification for those procedures. • When applicable, reports of research (quantitative or qualitative) that show that variations resulting from different delivery forms/orders do not alter the interpretations of results.

ADDITIONAL CONSIDERATIONS RELATED TO THE OVERALL PEER REVIEW PROCESS

Beyond the examples of evidence in the current guidance, experts noted some important aspects of the peer review process that may not require updating any aspect of the guidance itself, but nonetheless are crucial to supporting states in developing high-quality assessment systems, especially innovative ones.

Clarifying the Purpose of the Examples

Several experts noted a perception among states submitting to peer review that the "examples of evidence" are viewed as a checklist outlining what a state *must* submit, rather than as a set of illustrative types of evidence that a state might submit or not, depending on the appropriateness of the evidence for the given assessment system. Although the peer review guidance explicitly states that this is not the case in the paragraphs preceding the examples, by placing all of the examples in a list format, there is an implication that every example is (a) necessary and (b) weighted equally to all others. While it is the case that certain examples are going to be central to any successful submission, experts suggested that training around the peer review process for both states and peers should make it clear that the examples are just that—examples.

Providing Examples of Successful Submissions

An expert noted that states may find the process of submitting evidence for an innovative assessment system intimidating due to the existence of very few strong examples. Experts suggested that resources for submitters could include links to concrete examples of evidence from submissions that have passed peer review, with this being especially important for innovative/novel approaches. Building on the previous recommendation, one can see how the inclusion of strong example submissions might help states understand how the evidence associated with each critical element is consistent with the design of the assessment system. Particular emphasis was placed on examples of successful submissions with culturally responsive or anti-racist focus, given the rapidly evolving landscape of assessment methods in this area (e.g., Buzick et al., 2023).

Prioritizing Submission Coherence

Several of the psychometricians involved in the April 2024 meeting put forward innovative models in which a single summative determination was derived from multiple tests that varied in order, length, and/or content. It was noted by one expert that the text outside of the examples of evidence (such as in the *Assessments* section on pages 24–26) would benefit greatly from a paragraph encouraging states to consider coherence across all the separate elements of their systems. Although parts of the present report encourage states to only provide evidence for the scores relevant to the summative determination, it is crucial that states first consider how the individual pieces affect the whole. For instance, this could include whether scores are invariant to the order in which tests are taken, which item pools are included with each administration, variations in expectations for security, and the purpose of each individual administration in its relationship to the final determination. Including these considerations in a paragraph outside of the examples of evidence may help states in navigating the evidentiary needs in submitting their own innovative models.

POSSIBLE FUTURE DIRECTIONS

Experts participating in this project noted several important future directions. While they fall outside the scope of this report, each represents a potential direction for peer review for which the experts at the convening expressed enthusiasm.

Accounting for Testing Consequences

Experts noted that the role of testing consequences as an aspect of validity evidence is outlined in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), but is mentioned nowhere in the guidance. There was considerable enthusiasm for the prospect of incorporating validity evidence related to testing consequences as part of the peer review process, though some experts also warned that consequences are often dependent upon aspects of assessment policy outside the control of test developers—though this does not mean that developers should ignore the way that the assessment fits into the larger policy landscape. One possible route by which consequences could be addressed in peer review is the addition of a new Critical Element, accompanied by new examples.

Broadening Critical Element 4.2 – Fairness and Accessibility

One expert observed that the text of Critical Element 4.2 only requires fairness in design, development, and analysis. However, the joint standards include design, development, administration, and analysis. Expanding the text of the critical element to include fairness in administration would require a more substantial change to the guidance. It would also encourage states to provide evidence that their models (including innovative models with variations in order and timing of administration) have taken steps to ensure no groups (e.g., high mobility students) are disadvantaged.

Acknowledging the Native Language Provision in ESSA

The federal regulations for the *Every Student Succeeds Act* contain a specific provision that allows for those students in native language immersion programs to be assessed in the language of instruction (Section 200.6(j)). High quality native language instructional programs do not offer translated versions of English instruction, but instead teach rigorous, college- and career-ready standards grounded in the linguistic traditions of the specific native language. In the cases where a state elects to develop and administer a native language assessment for students receiving instruction in their native language, future iterations of the peer review guidance should provide examples of how states might demonstrate those assessments are reflective of the instructed content, while also providing for comparable inferences related to college and career readiness.

REFERENCES

Achieve (2019). *A framework to evaluate cognitive complexity in mathematics assessments.* https://www.achieve.org/files/Mathematics%20Cognitive%20Complexity%20Framework_Final_92619.pdf

AERA, APA, & NCME (2014). *Standard for educational and psychological testing.* American Educational Research Association. https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf

Buzick, H. M., Casabianca, J. M., & Gholson, M. L. (2023). Personalizing large-scale assessment in practice. *Educational Measurement: Issues and Practice, 42*(2), 5–11. https://doi.org/10.1111/emip.12551

Dadey, N. (2024, March 6). *Challenges for through-year and other innovative test designs*. Center for Assessment. https://www.nciea.org/blog/peer-review-and-innovative-test-design/

Every Student Succeeds Act, 20 U.S.C. § 6301 (2015). congress.gov/114/plaws/publ95/PLAW-114publ95.pdf

Forte, E. (2017). *Evaluating alignment in large-scale standards-based assessment systems.* Technical Issues in Large Scale Assessment State Collaborative on Assessments and Student Standards of the Council of Chief State School Officers. https://ccsso.org/sites/default/files/2018-07/TILSA%20Evaluating%20Alignment%20in%20Large-Scale%20Standards-Based%20Assessment%20Systems.pdf

Gong, B., & Patelis, T. (2016). *Guide to evaluating assessment using the CCSSO criteria for high quality assessment: Focus on test content.* Center for Assessment. https://www.nciea.org/wp-content/uploads/2021/11/CFA-Guide-FocusOnTestContent-R1_0.pdf

Office of Elementary and Secondary Education, Department of Education. (2016). *Every Student Succeeds—Innovative Assessment Demonstration Authority* (34 CFR Part 200) (Docket ID ED–2016–OESE–0047; RIN 1810–AB31). Federal Register, 81(236). https://www.govinfo.gov/content/pkg/FR-2016-12-08/pdf/2016-29126.pdf

U.S. Department of Education (2018). *A State's Guide to the U.S. Department of Education's Assessment Peer Review Process.* https://oese.ed.gov/files/2020/07/assessmentpeerreview.pdf

Enhancing Peer Review: Supporting Innovation in State Assessment Systems

Lyons ASSESSMENT CONSULTING

FORESIGHT LAW+POLICY