Evaluating Technical Quality in the Context of Assessment Innovation

Policy Implications from a Case Study

Susan Lyons, Sara Christopherson, Juliana Charles Brown,
Samuel Ihlenfeldt, Lillian Pace

Lyons, S., Christopherson, S., Brown, J.C., Ihlenfeldt, S., & Pace, L. (2025). *Evaluating Technical Quality in the Context of Assessment Innovation: Policy Implications from a Case Study*. Lyons Assessment Consulting, Wisconsin Center for Education Products and Services, KnowledgeWorks.

Table of Contents

Executive Summary

This report explores the evaluation of technical quality in the context of assessment innovation, where the design and use of assessments are intended to support deeper learning. We note that current existing federal peer review expectations are not well-suited for accommodating the design considerations and tradeoffs associated with innovative programs that intend to shift the purpose of statewide assessment beyond school identification and toward the transformation of teaching and learning. We propose a shift in federal assessment peer review processes to require the submission of a comprehensive validity argument, to more flexibly support states in gathering and submitting evidence related to the quality of the assessment system for serving its intended purposes.

To ground these ideas, we present findings from a case study of the Massachusetts Consortium for Innovative Education Assessment, which is piloting a curriculum-embedded, educator-developed performance assessment system known as the Portfolios of Performance. The Portfolios of Performance system is explicitly designed with students in mind, centering equity, authenticity and agency. The system was evaluated for alignment with academic content standards and score comparability with the state's current summative assessment, the Massachusetts Comprehensive Assessment System, to better understand the compatibility of performance-based systems with existing federal requirements. In doing so, the report proposes a pathway for performance assessments (and innovative assessment systems in general) to meet federal peer review requirements.

Findings indicate that the Portfolios of Performance assessment system showed strong potential to meet both existing federal expectations for alignment as well as more expansive alignment considerations such as the assessment's reflection of key instructional shifts embedded in the standards, including interdisciplinary connections and authentic engagement with disciplinary content. However, the assessment system has not yet met the necessary thresholds for score comparability. Recommendations for strengthening the system include increasing the number of tasks, refining task quality and enhancing educator scoring reliability through targeted training and calibration.

More broadly, this report argues that the evaluation of assessment quality, particularly for models intended to serve as tools for instructional improvement, should be structured around a well-reasoned and coherent validity argument. Such an argument could include a clear theory of action, evidence of theoretical coherence, expanded evidence of alignment, appropriate methods for evaluating comparability and ongoing attention to implementation and systemic impacts.

We close with policy recommendations for state and federal leaders, including the need to:

- Strengthen support for the development and continuous improvement of innovative assessment models
- Create the conditions to improve performance assessment quality within the current federal framework
- Reorient the federal assessment paradigm to enable and encourage the assessment of deeper learning

Taken together, these recommendations offer a pathway for building assessment systems that are not only technically sound but also better support the goals of equity, instructional relevance and deeper learning.

Introduction

Deeper learning encompasses a set of education goals and priorities that have long been valued and have appeared under various names throughout the history of education within the field of curriculum and instruction. These high-level goals include educational outcomes that are cognitive, interpersonal and dispositional. The term "deeper learning" was introduced by The William and Flora Hewlett Foundation in 2010 and helped focus national attention on key underlying concepts that were integral to the 2000-era educational reform movement and are now emphasized in contemporary academic content standards, such as the Common Core State Standards, Next Generation Science Standards and similar frameworks. These standards prioritize not only core academic knowledge but also essential skills like critical thinking, creativity, collaboration and communication. In the context of educational assessment, the term “innovative assessments” typically refers to assessment formats other than highly standardized, on-demand formats such as multiple choice. Although the findings presented in this report can be applied to many forms of innovative assessments, the central focus relates to the use of instructionally embedded authentic performance assessments. With an emphasis on real-world tasks and problem solving, authentic performance assessments closely reflect the goals of the deeper learning movement. Curriculum-embedded, performance-based measures of student achievement are considered powerful tools to promote equity in the classroom (Darling-Hammond, 2017) and, in practice, have the potential to promote more equitable learning environments by building on students’ funds of knowledge and lived experiences (e.g., Diaz-Bilello & Pierre-Louis, 2021).

Those who espouse the adoption and use of performance assessment systems often do so in service of a theory of action in support of deeper learning goals (see, for example, Darling-Hammond et al., 2010). There are multiple ways that the use of performance assessment is thought to improve student outcomes. For example:

- As high-quality learning experiences in and of themselves
- As a signaling mechanism to model the kinds of instructional tasks that represent the deeper learning goals of the content standards
- As a feedback mechanism to provide insight into student thinking and work processes (Student et al., 2023)

Each of these mechanisms for improving student learning through performance assessment is closely related to the teaching and learning processes. In contrast, the broader theory of action for standardized, test-based accountability relies less on the educative nature of the assessment and more on the monitoring, identification and public reporting of student achievement and subsequent state interventions. These two lenses lead to different priorities related to the properties of the test and, therefore, imply the need for distinct evidence of technical quality when evaluating the suitability of an assessment for a particular context.

We propose considerations for how technical quality may be evaluated in the context of assessment innovation, where the test itself is intended to be a tool for transformation (i.e., leading to improvements in teaching and learning). We ground these proposed considerations in findings from a case study evaluation of a performance assessment system. The Massachusetts Consortium for Innovative Education Assessment (MCIEA) Portfolios of Performance (PoP) systems in grade three mathematics and grades six, seven and eight English language arts were evaluated against the current federal framework for technical quality, specifically as they relate to alignment and score comparability, to understand where the existing federal framework could be expanded or improved to better accommodate innovative assessment systems that have a closer connection to teaching and learning than traditional state summative models. We close with recommendations for state leaders and federal policymakers to support a paradigm shift in statewide assessment systems toward a model that creates space for and encourages innovative approaches that emphasize transformation, alongside traditional uses focused on identification.

Prioritizing Validity Argumentation in Peer Review

Since the passage of the No Child Left Behind Act (NCLB) in 2001, there has been no shortage of public criticism of statewide standardized tests (Richardson, 2017). Though the criticisms vary by context, many of them have their root in the lack of value provided by the tests for both educators and learners. Secure test administration takes substantial time away from instruction, and the resulting test scores provide little actionable information for any given classroom or learner. State leaders often contend that improved communication with the public could help clarify that statewide summative tests are designed to allow districts, states and policymakers to identify schools in need of support and are not intended to serve as tools for teaching and learning. A prevailing belief in the broader field of assessment is that assessments designed for accountability cannot and should not be integrated into instructional practices (Figure 1). While this may currently be true operationally, it is likely dissatisfying to those who are subjected to lengthy annual testing schedules with limited perceptible benefit for their students.

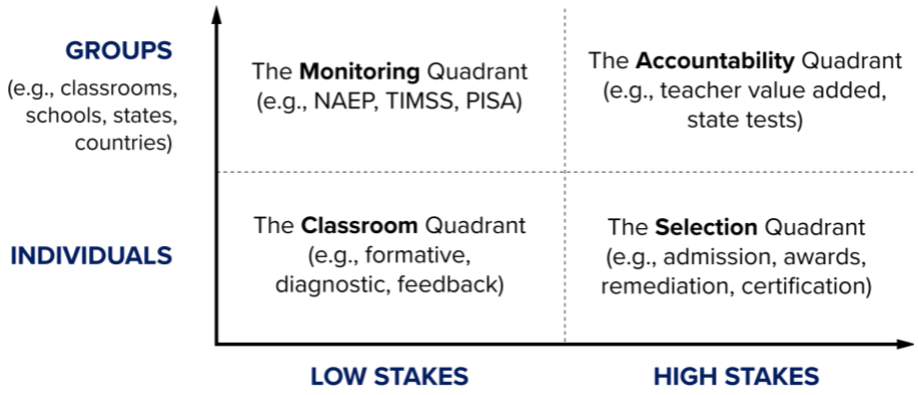

Figure 1. *Four Purposes of Educational Tests (adapted from Ho, 2022)*

In light of these concerns, there has been renewed interest in adopting a range of innovative designs, such as through-year testing, curriculum-embedded tasks and performance assessments that aim to produce more instructionally useful and locally relevant information. While innovative models differ in approach, many share a commitment to expanding the purpose of large-scale assessment beyond the accountability quadrant (Figure 1). The focus of this report is performance assessments, using MCIEA's PoP system as a case study. Montana's current through-year model provides another example of how innovative models might be designed with consideration of peer review requirements. The Montana Aligned to Standards Through-Year Assessment (MAST) is built on a theory of action for informing instructional adjustments with more relevant and timely assessment information (MAST Task Force, 2022). By setting aside claims and evidence related to supporting teaching and learning and focusing solely on claims and evidence related to accountability, the Montana model shows a potential approach for an innovative assessment system, when designed to inform and improve learning, to meet current federal requirements for technical quality. Despite this example, many experts argue that the federal Title I peer review process remains poorly suited to support more transformative assessment models (Ihlenfeldt et al., 2024), especially those centered on instructional relevance and local context. Revising peer

review guidance to more flexibly value claims and evidence tied to teaching and learning within a validity argument could more strongly connect federal expectations with the intended test purposes driving many innovations.

Developing an argument-based approach to peer review would more prominently underscore the reality that all assessment design decisions come with inherent tradeoffs. For example, in federal peer review guidance, an orientation around standardization of administration (Critical Element 2.3-2.4) and scoring (Critical Element 4.4) is presented as a given, rather than as an intentional choice (U.S. Department of Education, 2018). While standardization is efficient for gathering evidence of alignment and score comparability, it is not the only possible approach. It comes with its own set of tradeoffs, such as limiting instructional utility.

Modern validity theory positions assessment system validation as an argument consisting of a series of logical chains of reasoning supported by theory and evidence (Kane, 1992; 2013). Determining how to structure that network of inferences and assumptions for transparent evaluation is, however, far from intuitive. Claims can rarely be fully substantiated beyond scrutiny and tend to require an amassing of evidence, showing that the strength of the claim is directly reliant on the strength of the supporting theory and evidence. As a solution, measurement theorists (Kane, 2004; Mislevy et al., 2003) have turned to Toulmin's model (Figure 2) as a structure for validity arguments. Toulmin's model breaks down reasoning into key components: a claim supported by data and connected through a warrant and alternative explanations. Additional elements like backing and rebuttals help clarify the strength and limits of the argument. The model emphasizes real-world reasoning over formal logic; thus allowing assessment designers to argue for the intended uses of an assessment while still acknowledging potential tradeoffs attributable to the assessment design.

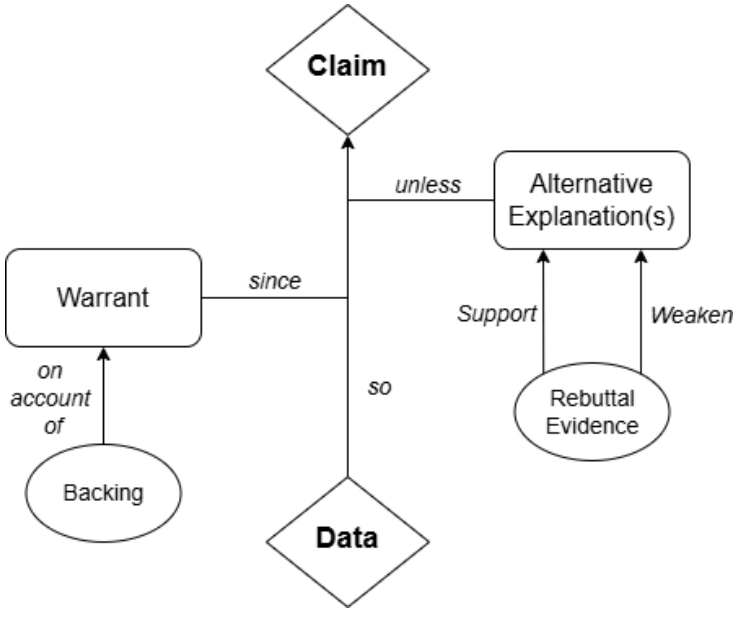

Figure 2. *Adapted from Toulmin's (1958) structure for arguments (cited in Mislevy et al., 2003)*

Current peer review guidance lays out expectations for validity evidence according to the categories identified in the *Standards for Educational and Psychological Testing* (2014): evidence based on test content (Critical Element 3.1), response/cognitive processes (Critical Element 3.2), internal structure (Critical Element 3.3) and relations to other variables (Critical Element 3.4). Evidence related to the consequences of testing, included in the *Standards for Educational and Psychological Testing*, is not included in the current peer review guidance. Although intended to prompt the submission of a validity argument consistent with professional standards, peer review expectations are commonly approached as a checklist for compliance. Simply put, the current peer review framework does not require an argument structure for the evidence submitted. This may unintentionally limit presenting a range of evidence to support the intended score claims and uses as allowed by the flexibility of the framework.

We join other measurement experts in suggesting that the guidance be clarified and expanded to prioritize coherent validity arguments that require a thoughtful structuring of evidence into a chain of reasoning that supports the intended interpretations and uses of the test scores and reflects the purpose of the assessment (Ihlenfeldt et al., 2024). In the sections that follow, we elaborate on a framing of peer review that emphasizes a validity argument to defend the appropriateness of an assessment for supporting its intended purposes and uses. Incorporating the practice of validity argumentation into the technical quality review of statewide summative assessments could create space for states to defend their design decisions and tradeoffs when considering the intended purposes of the tests. In cases where more transformative purposes related to teaching and learning are at the center of the assessment design, technical quality evidence would necessarily look different. For example, alignment criteria might expand to emphasize pedagogical coherence, and likewise, methods for gathering evidence of score comparability could differ from the more traditional reliance on standardization.

Presenting a Validity Argument in the Context of Innovation

A well-structured validity argument creates transparency for users to evaluate the sufficiency and coherence of evidence supporting the intended score interpretations and uses. Evaluating an assessment's technical quality by considering the strength of an argument about the appropriateness of the test design and evidence for supporting the test uses simultaneously creates a higher standard of rigor for test vendors while opening doors for additional forms of evidence not specified in the current peer review system. We elaborate on examples of additional evidence that could be particularly relevant in the context of evaluating the quality of innovative assessments in the sections below. This includes:

1. Developing a theory of action
2. Prioritizing theoretical coherence
3. Expanding evidence of alignment
4. Gathering evidence of comparability
5. Monitoring the integrity of implementation
6. Evaluating systemic impacts

Developing a Theory of Action

Federal peer review guidance expects states to submit a statement of the purpose(s) of an assessment and the intended interpretations and uses of results (Critical Element 2.1, U.S. Department of Education, 2018). Recently, experts have called for the inclusion of a test rationale, such as a theory of action, within federal peer review guidance to link purpose with use in support of a more transparent and intentional assessment design, especially in the context of innovative systems (Ihlenfeldt et al., 2024). Bennett (2010) argued that the development of a theory of action is central to the design of an assessment and is fundamentally incorporated within a validity argument. He proposed that the theory of action for assessment systems could include:

- "the intended effects of the assessment system;
- the components of the assessment system and a logical and coherent rationale for each component, including backing for that rationale in research and theory;
- the interpretive claims that will be made from assessment results;
- the action mechanisms designed to cause the intended effects; and
- potential unintended negative effects and what will be done to mitigate them" (Bennett, 2010, p. 71).

These arguments echo Messick's (1994) assertion that validity evidence should support not just the direct interpretation of scores but also the inferences that result in the applied decisions made and actions taken based on test scores.

A theory of action outlines the key components of an assessment system and makes explicit how those components are connected. It is more than just a planning tool; a theory of action is a framework grounded in research and logic that guides assessment design by making clear how the system's components are intended to work together to achieve specific goals. A theory of action is a framework grounded in research and logic that guides assessment design by making clear how the system's components are intended to work together to achieve specific goals. It should articulate both the components of the system and the logical connections among them. Critically, the theory of action lays out the mechanisms through which the system is expected to drive change (Marion, 2010). Lane (2014) proposed that a theory of action becomes an organizing structure for validation, especially for evaluating the consequences (both intended and unintended) of an assessment system designed to impact teaching and learning. More recently, researchers have reiterated this, suggesting that theories of action are a way to develop culturally relevant assessment systems (Englert & Shultz, 2025) or address criticisms of testing in a shifting landscape (Markle, 2024).

In practice, theories of action have served as essential foundations for the design and implementation of innovative assessment systems. Lyons et al. (2017), for instance, argued that the validity of the New Hampshire Performance Assessment of Competency Education (PACE) should be understood as an evidence-based argument anchored in its theory of action rather than treated as a binary judgment. Similarly, the Montana Office of Public Instruction (2022) employed

the theory of action guiding the MAST pilot to determine the critical resources and types of evidence needed to support both system development and implementation. For assessment systems intended to influence instructional practice and learning environments, articulating a clear theory of action is especially critical, as the types of evidence of technical quality will expand beyond and could be different from the scope of evidence specified in the current federal peer review guidance (e.g., coherence with curriculum and instruction, utility of score reports). The types of evidence of technical quality will expand beyond and could be different from the scope of evidence specified in the current federal peer review guidance (e.g., coherence with curriculum and instruction, utility of score reports).

Prioritizing Theoretical Coherence of Assessment Design

In the seminal National Research Council report *Knowing What Students Know*, Pellegrino et al. (2001) argue that all assessment designs rest on three interrelated processes: cognition, observation and interpretation, illustrated in Figure 3.

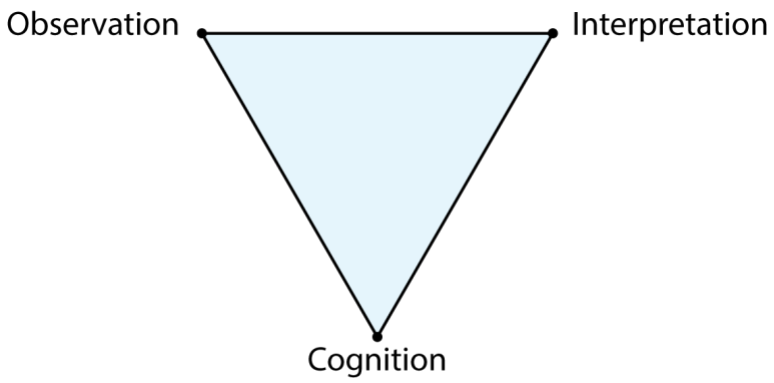

Figure 3. *The Assessment Triangle (adapted from Pellegrino et al., 2001)*

The cognition vertex represents an understanding of how students think and learn relative to the targeted domain. Traditional, on-demand assessments tend to define greater achievement by the ability to answer more difficult questions. However, observations of student learning that are based on the ability to answer difficult questions do not link tightly with a contemporary understanding of how students learn and develop disciplinary knowledge and epistemic practices. Instead, learning is generally understood as a process that follows non-linear and potentially interacting sequences (National Academies of Sciences, Engineering, and Medicine, 2007, p. 221; Songer and Gotwals, 2012) and is infinitely variable across learners (Pape, 2018). Motivation, identity and culture are closely intertwined with cognition and learning (National Academies of Sciences, Medicine, Division of Behavioral, Social Sciences, 2018). Curriculum-embedded performance assessments and other types of innovative assessments may be better suited to provide opportunities that reflect current perspectives on learning. For example, they may include tasks that can be responsive to factors related to the student context (e.g., socioculturally relevant tasks) and allow students multiple inroads to demonstrate learning through a variety of pathways. Observations of student learning elicited from these types of tasks will yield different interpretations than tasks designed with an emphasis on standardization.

Tasks that closely reflect our best understanding of teaching and learning are more likely to lead to usable insights for students and educators as they adjust their learning and instruction. Therefore, one dimension of an assessment's technical quality can be viewed in terms of the theoretical coherence across the three vertices of the assessment triangle, consistent with a holistic perspective of alignment. In other words, how we observe and interpret student learning should be closely linked to the ways we understand learning as occurring. For example, if we understand learning as a stepwise, linear progression, then we might observe evidence of learning using a series of sequenced tasks and interpret greater achievement to be represented by a student moving further along the sequence of tasks. Alternatively, if we understand learning as a network of ideas that gets revised and elaborated as learning progresses, then we might observe evidence of learning using a variety of tasks that would provide evidence that a student could use multiple strategies to solve a problem. In this case, we might interpret greater achievement to be represented by a student's ability to demonstrate the use of a range of strategies.

In a coherent system, learning goals that emphasize deeper learning are assessed using tasks that provide opportunities for students to demonstrate principles of deeper learning. Innovative assessment programs should clearly detail how the program design reflects both the program's theory of learning and the interpretations it intends to support. Compared with traditional, on-demand assessments, innovative assessment programs designed to support teaching and learning may provide opportunities that more closely reflect the model of cognition underlying the academic content standards, and this should be made explicit within the assessment's validity argument.

Expanding Evidence of Alignment

The premise of standards-based education is that setting high expectations for all students can advance positive change through systemic coherence. When the shift toward standards-based education in the 1990s started, the idea of alignment was not clearly articulated in the field. While the importance of systemic coherence was recognized, the content alignment relationship between standards and assessments was of particular interest because standards and assessments were being employed as high-leverage policy elements intended to influence teaching and learning in positive ways.

With a focus on the policy elements of standards and assessments, a literature review suggested five categories of key alignment criteria: those related to test content, articulation across grades and ages, equity and fairness, pedagogical implications and system applicability (Webb, 1997). The National Institute for Science Education, funded by the National Science Foundation and in collaboration with the Council of Chief State School Officers, convened a panel of experts to further refine these key alignment considerations. It was broadly recognized that a fragmented system would send mixed messages. Although not intended as a definitive list, the purpose of this work was to help the field think more clearly about what meeting the systemic reform goals of a standards-based system would entail. The resulting research monograph (Webb, 1997) described 12 alignment criteria organized into the five categories identified through the literature review

(Figure 4). These categories and criteria were developed with a perspective of a broad system of assessments and not with a focus just on statewide summative tests. However, because mandated statewide standards and assessments were being used to make consequential decisions, there was particular attention given to the importance of alignment in this context.

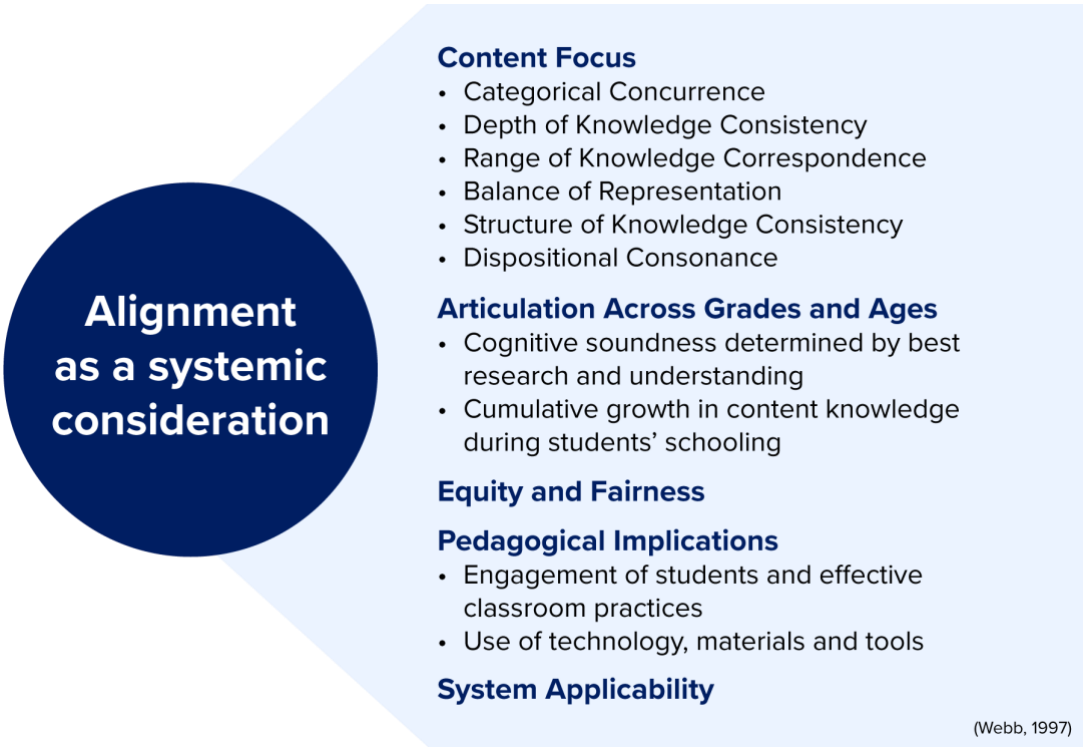

Figure 4. *Alignment criteria proposed in Webb's 1997 research monograph*

In addition to defining alignment criteria, there was interest in developing an evaluation framework and methodology that could be used to evaluate the alignment of academic expectations and assessments in consistent, fair, transparent and reproducible ways. Based on the purpose of the work, to operationalize alignment for statewide summative assessments, there was a practical decision to limit the scope of the work specifically to the criteria within the content focus category (Figure 4). Expert panels trialed processes, and their feedback was used to refine the approach, resulting in the further omission of two of the six criteria within the content focus category because, at that point, statewide assessments were not attending to the structure of knowledge nor student disposition (Webb, 1999).

Today, the concept of alignment in the context of education is often used to refer specifically to content alignment between different components of educational materials, including between standards (or other assessment targets) and assessments. Within the field of educational assessment, alignment typically refers to the relationship between a set of academic standards or assessment targets derived from the standards (the referent) and a test (the comparand) (Fulmer, 2011; Porter, 2002; Webb, 1997). It is broadly recognized that alignment is the anchoring concept at the heart of the academic standards movement (Polikoff, 2020) and that content alignment evidence is central to a validity argument for an assessment (American Educational Research

Association, APA & National Council on Measurements in Education, 2014). However, alignment is still a relatively new field. The *Standards for Education and Psychological Testing* did not include references to content alignment until the 2014 edition.

Although the term "alignment" was first included in the 2014 *Standards for Education and Psychological Testing*, similar ideas had long existed within the field of measurement and were addressed within the spheres of "content validity" and "consequential validity" (Traynor & Christopherson, 2024). Professional standards continue to affirm the importance of validity evidence based on test content, including evidence of content alignment, to a comprehensive validity argument.

Federal expectations for alignment are often distilled into the phrase "depth and breadth" or, sometimes, "depth, breadth, and balance." An assessment is expected to address some breadth of the standards (or other measurement targets). This expectation extends from earlier "content validity" expectations for a test to demonstrate "content representativeness" (Traynor & Christopherson, 2024). "Depth" refers to cognitive demands such as conceptual understanding, critical thinking and problem solving. Because these types of cognitive demands were intended to be communicated through the language of the standards and were central to the deeper learning goals of standards-based reform, a coherent system required this idea of "depth" to be identifiable through content analyses and reflected in the content of the curriculum and assessments. Balance refers to the extent of emphasis on any particular assessment target. In an aligned system, any emphasis on particular assessment targets should be purposeful.

Depth, breadth and balance are important content alignment considerations. However, this subset of alignment criteria is insufficient to address the types of systemic changes intended by a standards-based system, including, importantly, the central purpose of promoting improved teaching and learning. While cognitive goals (i.e., critical thinking, problem solving, etc.) were and continue to be highly valued within academic standards, additional goals include productive contribution to collaborative work, creativity, cultural relevance, student agency, self-directed learning and others. Affective goals could include the development of positive attitudes and beliefs about learning, a sense of self-efficacy and others. These components are emphasized in Webb's (1997) attention to alignment criteria such as dispositional consonance and the pedagogical implications of assessments. In an aligned system, the instructional practices and classroom learning contexts implied by the standards should be reflected in the assessments. Overall, the goal for alignment is a coherent system where all the policy elements and component parts are working together toward the common goal of improved student outcomes supported by effective learning experiences.

Given this history and fuller context, the current federal peer review expectations emphasize a limited set of specific criteria related to alignment between standards and assessments. For purposes of submission of evidence to federal peer review, states often defer to this set of minimum expectations. While the criteria of depth, breadth and balance are important, we argue that other criteria, such as those initially proposed by Webb (1997) and those valued and prioritized by districts and states, also merit attention. In the context of a validity argument, a vendor and state could support the strength of their assessment model by incorporating an evaluation of additional alignment criteria, along with those related to content focus.

Gathering Evidence of Comparability

Statewide standardized achievement tests typically support claims of score comparability across students by standardizing the items students respond to and the administration conditions. While this is efficient, it requires strict adherence to a single set of equated test forms (or test events in the case of computer adaptive testing) as well as secure and structured administration procedures. These design decisions have led to testing occasions that feel disconnected from and disruptive to classroom learning. Assessments that prioritize deeper connections to teaching and learning processes are likely to lead to a different set of test design decisions, allowing for local or even individual flexibility in tasks and administrations that emphasize connections to instruction and sociocultural context. These decisions introduce tradeoffs for both the strength of the comparability claims (e.g., strict scale score comparability may never be a realistic target) and the methods for gathering evidence of comparability.

Peer review guidance that explicitly values structured and context-specific validity arguments would give states space to discuss and defend their design decisions in light of the intended assessment purposes, and their methods for gathering evidence of score comparability at the level necessary to support the intended claims. For innovative systems, especially those designed to produce more instructionally relevant information, establishing comparability at the scale-score level may not be appropriate. For statewide testing programs, the level of comparability needed to support identification of schools rests at the proficiency determination, not the scale score. If the determination of student proficiency is sufficiently comparable across students and schools, the assessment meets the current federal requirements for use within the school accountability systems. This introduces a wide range of design possibilities, including those used to produce individually meaningful information for teaching and learning (e.g., exposing student thinking and processes), while simultaneously serving the policy function of school identification. For socio-culturally relevant systems, such as those that leverage authentic performance assessments, large-scale testing programs have largely relied on methods of social moderation to produce comparable achievement level determinations (e.g., Advanced Placement, International Baccalaureate). These approaches are not yet standard in most statewide summative systems, but could be fully supported in a peer review process that prioritizes an argument-based approach to validation.

Monitoring the Integrity of Implementation

Innovative assessment systems that attempt to tie summative assessment more closely to teaching and learning processes do not come without additional burdens on systems, schools and educators. For example, New Hampshire's innovative testing pilot, PACE, which worked to link summative assessment more directly to the local curriculum, came with incredibly high resource and time burdens related to professional learning and implementation monitoring (Troppe et al., 2023). Given this, a comprehensive validity argument for innovative assessment systems should include evidence of the elements essential to successful program interpretation and implementation across contexts.

LeMahieu (2011) distinguishes the concept of *integrity of implementation* from the more commonly used, *fidelity of implementation*, to "allow for programmatic expression in a manner that remains true to essential empirically warranted ideas while being responsive to varied conditions and contexts" of implementation. For example, in the context of an assessment system that prioritizes performance tasks, a task that requires an authentic audience will likely need to substantively vary across contexts to serve its intended functions. Organizing a presentation to the local urban planning board might work well in one community, while drafting a letter about a United Farm Worker campaign might serve better in a different context. This kind of local adaptation of the assessment program to maintain its essential qualities takes planning, time and significant investment in local capacity. A perfectly designed assessment model can fail in poor implementation, and all reasonable evaluations of quality in the context of this innovation must pay particular attention to those factors that contribute to the integrity of implementation across contexts.

Evaluating the Systemic Impacts

The underlying rationale for the entire standards-based education reform movement is the goal of systemic school improvement. A validity argument that values evidence related to testing consequences will require an explicit articulation of the intended systemic impacts. Innovative testing programs might consider key questions that include, but are not limited to:

- Are students having meaningful assessment experiences?
- Are educators getting the information they need to adjust practice?
- Is instruction changing to reflect deeper learning priorities?
- Are parents, caregivers and policymakers receiving accessible and actionable information about student learning?
- Are state leaders able to effectively identify low-performing schools?
- Is student learning improving?

Messick (1992) contends that the validity argument for a performance assessment should incorporate both the intended instructional benefits and any potential negative consequences. Unsurprisingly, the impacts of an assessment program, both intended and unintended, have long been recognized as central to the evaluation of test quality.

For example, within the *Standards for Psychological and Educational Testing* (2014), Standard 1.6 states:

> "When a test use is recommended on the grounds that testing or the testing program itself will result in some indirect benefit, in addition to the utility of information from the interpretation of the test scores themselves, the recommender should make explicit the rationale for anticipating the indirect benefit. Logical or theoretical arguments and empirical evidence for the indirect benefit should be provided. Appropriate weight should be given to any contradictory findings in the scientific literature, including findings suggesting important indirect outcomes other than those predicted" (p. 24).

This means that theoretical and empirical evidence of the expected benefits of a testing program should be furnished as part of the ongoing evaluation and maintenance of an operational assessment. Notably, if there is evidence that the testing program does not result in the intended benefit, that should be brought to bear. All testing programs, particularly those operating in the context of assessment innovation, should include a research program that seeks to collect validity evidence related to testing consequences, both intended and unintended, on student learning as well as on the school and state systems in which the assessments are operating.

Lane (2014) argues that a theory of action, discussed above, can provide the structure needed to identify, anticipate and evaluate both the positive and negative outcomes of an assessment system. Lane illustrated this approach through the case of the Maryland School Performance Assessment Program (MSPAP), a performance assessment that was discontinued, like many similar assessments, after the passage of NCLB, because it was considered incompatible with the new federal expectations. In the analysis, Lane outlined five core interpretive claims tied to the program's intended outcomes, along with corresponding sources of evidence to support each claim. Importantly, she also identified several potential negative consequences. Together, the positive and negative systemic impacts drove the collection of validity evidence for the MSPAP.

Ultimately, to keep pace with modern measurement theory and practice, federal peer review should signal that the process values and prioritizes the strength of an assessment's comprehensive validity argument, including evidence related to positive and negative consequences. Validity evidence related to testing consequences is currently the only source of validity evidence that is fully omitted from the federal peer review guidelines. While evidence related to testing consequences is important for all assessment programs, this source of validity evidence becomes particularly critical for states seeking to design, implement and evaluate innovative systems of assessment that are intended to transform teaching and learning.

A well-structured validity argument that includes evidence and theory related to the elements discussed above will help stakeholders clarify the intended purposes of the assessment, understand design choices and tradeoffs and evaluate the strength of the evidentiary support for the testing program claims.

The following sections of this paper introduce the current policy context for performance assessment and peer review in the United States and dive deeply into a case study that evaluates a particular performance assessment system in light of a portion of the current federal framework for assessment technical quality. We conclude with policy recommendations for state and federal leaders that are intended to shift the narrative and practice around gathering evidence of technical quality toward a more comprehensive and structured approach, such as what we have argued for here.

Current Policy Context for Assessment Innovation

Many states have a longstanding interest in performance assessments and have supported efforts to increase local adoption and use of these systems (e.g., Colorado, North Carolina, Virginia). Despite this interest and support, curriculum-embedded, authentic performance assessments have not been successfully adopted and scaled into statewide summative testing programs. Some states were beginning to use performance assessments at scale in the 1990s (e.g., Kentucky, Vermont, California, Maryland), but that work came to an end with the testing requirements introduced in the 2001 NCLB legislation. These have not been authentically and successfully reincorporated into statewide summative assessments since NCLB. States report that they hesitate to try performance assessments due to the administrative costs as well as concerns that they are incompatible with expectations as articulated in the federal assessment peer review guidance.

Recognizing risks of constraining innovation within statewide summative testing, policymakers successfully advocated for the inclusion of the Innovative Assessment Demonstration Authority (IADA) in the Every Student Succeeds Act (ESSA) to help promote the development and use of high-quality, innovative assessments. This authority permits a state educational agency to develop and pilot innovative assessments in a subset of districts before implementing those assessments statewide. Despite the program's high level of interest, the United States Department of Education (USED) has struggled to enlist and maintain state participation. Only six states have been awarded an IADA since 2018, and only three states remain active in the program. At the end of 2023, USED announced a rolling application process in which states have two chances to apply annually; since then, no states have applied.

After receiving informal feedback that states were discouraged from applying due to perceived barriers from federal requirements, USED released a request for information to solicit formal input (United States Department of Education, 2023). The feedback received emphasized state concerns about meeting score comparability and alignment requirements as stated in federal statute (e.g., The Education Trust, 2023). New guidance released by USED in the fall of 2023 included clarifications related to the expectations for demonstrating comparability for IADA approval.

Along with the new guidance, USED also lifted the cap on IADA and prioritized funding for states seeking to apply for IADA in the Competitive Grants for State Assessment (CGSA) competition in hopes of encouraging more states to participate (Applications for New Awards; Competitive Grants for State Assessments Program, 2024).

In the following section, we present a case study: an evaluation of a pilot performance assessment system in the context of today's expectations under ESSA. Our study works to identify specific pathways and barriers that federal requirements may present based on analyses of an actual innovative assessment system. Our research focuses on the topics of alignment and score comparability because those issues of technical quality present two of the most cited barriers for states designing innovative measures of student achievement to implement within their state accountability systems.

Massachusetts Consortium for Innovative Education Assessment Case Study

The Massachusetts Consortium for Innovative Education Assessment (MCIEA) was established in 2016 to create a model for a more equitable assessment and accountability system in the state of Massachusetts. MCIEA seeks to develop a viable measure of student achievement that leverages educator-generated, curriculum-embedded performance assessments. In addition to measuring student achievement of academic expectations as defined in the grade-level standards, these assessments are designed with a focus on equity, authenticity and student agency. During the 2023–2024 school year, four MCIEA districts engaged teams of educators to participate in a pilot that aimed to generate proficiency determinations using portfolios of student work from performance assessments in English language arts and mathematics.

The MCIEA case study presented a unique research opportunity because the consortium is not operated by a state agency and therefore does not have the authority to waive participation in the statewide assessment, the Massachusetts Comprehensive Assessment System (MCAS). Consequently, all participating students have summative annual determinations from both assessment systems (i.e., MCIEA's Portfolios of Performance [PoP] and MCAS), which allowed for a thorough exploration of comparability.

Research Goals

To guide our work, we used the following two research questions related specifically to the compatibility of MCIEA's performance-based measure of student achievement and federal requirements for Innovative Assessment Demonstration Authority (IADA) application under the Every Student Succeeds Act (ESSA):

1. To what degree are the federal statutory and regulatory requirements related to **alignment** with the challenging state academic content standards (as defined in §200.105(b)(2)(i) and 81 FR 88940 (Dec. 8, 2016)) compatible with curriculum-embedded performance assessment systems such as MCIEA's Portfolios of Performance?
2. To what degree are the statutory and regulatory requirements related to **score comparability** with the current statewide summative assessment (as defined in §200.105(b)(4)(i)(A) and 81 FR 88940 (Dec. 8, 2016)) compatible with curriculum-embedded performance assessment systems such as MCIEA's Portfolios of Performance?

Findings from these first two research questions helped answer two additional research questions related to policy implications and future system design recommendations:

3. Based on the findings of this study, what federal policy adjustments would better encourage and accommodate the use of innovative accountability measures such as performance assessment systems?
4. Based on the findings of this study, what recommendations can we offer for states interested in designing or refining innovative measures of student achievement in a way that is compatible with existing federal requirements?

Alignment Methods

Evidence of content alignment is a critical component of a validity argument for an assessment. The alignment approach used in this study was grounded in a process known as the "Webb alignment method." An initial outline of the methodology was described in detail in a research monograph, *Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education* (Webb, 1997) and has been refined and improved over the years, yielding a flexible, adaptable, effective and efficient analytical approach. Some version of this alignment methodology has been used to analyze curriculum standards and assessments in nearly all states to satisfy or to prepare to satisfy Title I compliance as required by USED. The alignment study design used for the analysis of MCIEA's PoP assessment was structured to yield evidence that would address the existing alignment requirements as stated in federal statute, with a focus on depth, breadth and balance. Additionally, reviewers evaluated tasks according to the criteria of equity, authenticity and agency as defined by the assessment program, based on a content analysis of each task along with other available supporting materials. These additional criteria fall outside of federal requirements for alignment but were important to the context of MCIEA's performance assessment program, and link to the fuller conceptualization of alignment described earlier in this report. For example, just as the standards are intended to promote student agency in the classroom, so should these pedagogical implications carry through to an assessment in an aligned system.
The alignment analysis detailed in this report was completed through the WebbAlign program, which operates out of the Wisconsin Center for Education Products and Services (WCEPS), a nonprofit organization affiliated with the University of Wisconsin-Madison. The WebbAlign program extends and expands the work of Dr. Norman Webb, Senior Research Scientist Emeritus of the Wisconsin Center for Education Research. If applied as intended, the Webb methodology expects a study design to be tailored to reflect the context of an assessment, taking the assessment framework (i.e., blueprint and other supporting documentation) into consideration in the study design. The WebbAlign program is characterized by setting clearly defined and agreed-upon alignment criteria in advance of data collection. These criteria should reflect the claims that are intended to be made based on student scores. Agreed-upon levels of acceptability for each criterion are also set in advance of data collection. These thresholds indicate what stakeholders consider "good enough" to meet each criterion. Stakeholders may choose to set absolute minimum thresholds or they may set higher thresholds, according to defined expectations. Decisions should be grounded in reasonable rationales so that the resulting evidence has the potential to support the intended measurement claims.

For example, consider an assessment that is intended to measure student proficiency as relates to grade five mathematics as defined by a state's academic standards. Even a summative assessment typically does not address every single grade-level standard. Instead, it samples knowledge and skills from across the full set of standards. What breadth of sampling would yield reasonably adequate evidence of what students know and can do as relates to the expectations within the grade five mathematics standards? The alignment criterion of Range of Knowledge Correspondence is used to judge whether a comparable span (breadth) of knowledge expected of students by a conceptual category (content domain or reporting category) is the same as, or corresponds to, the span of knowledge that students need to correctly answer the set of items or successfully engage with the set of tasks on an assessment. The minimum threshold originally proposed for this criterion was that an assessment should address at least 50% of the targets within each content domain included for assessment. This minimum threshold is based on the logical argument that to be able to determine if a student is proficient relative to an overall conceptual category, students' knowledge should be tested on content from over half of the of knowledge that is detailed in the standards (or other learning expectations) within that conceptual category. Particular circumstances may result in different decisions about appropriate thresholds. For example, stakeholders may expect sampling from most learning expectations within a domain rather than just more than 50%. For testing purposes, a system may reorganize standards from different content domains or organize standards to place greater value or emphasis on some standards over others. Multiple factors, including the specificity of the intended claims and the standards and assessment structures, need to be considered to determine the set of criteria to use as well as the appropriate threshold for acceptability for each criterion.

The IADA expectations for alignment evidence are the same as those that are expected for submission to federal assessment peer review. In the context of assessment peer review, test-event-level Webb alignment analyses typically address, at minimum, four specific content-focused criteria related to the agreement between the expectations within the standards and the demands of the items/tasks within the assessments, as shown in Table 1. See Appendix A for more detail.

Table 1. Four Content-focused Criteria Commonly Used to Evaluate Content Alignment of Assessments with Standards (Webb, 1997)

Criterion	Description of Criterion	Proposed Minimum Cutoffs* for Acceptable Alignment
Categorical Concurrence	The same or consistent reporting categories, or content domains, appear in both standards and assessment documents. Test forms have the potential to yield sufficient evidence to make inferences about student proficiency as relates to each reporting category.	Unless otherwise specified by test design, and in the context of assessment peer review, a test form is often expected to have at least six items measuring content from a reporting category.
Depth of Knowledge (DOK) Consistency	The assessment elicits work that is as cognitively demanding as the expectations in the corresponding assessment targets.	Unless otherwise specified by test design, at least 50% of the assessment items corresponding to standards within a domain are at (or above, although not common) the DOK level of the corresponding standards.
Range of Knowledge Correspondence	A comparable span of knowledge expected of students by a reporting category is the same as, or corresponds to, the span of knowledge that students need to correctly answer or successfully complete the assessment items/tasks.	Unless otherwise specified by test design, at least 50% of the standards for a reporting category are addressed by at least one related assessment item or student interaction.
Balance of Representation	A single assessed standard should not be (unintentionally) overrepresented on a test form. This criterion is used to indicate the degree to which one standard is given more emphasis in the assessment than another.	Unless otherwise specified by test design, an index value of 0.7 or higher is obtained, based on the difference in the proportion of standards addressed by items and the proportion of items corresponding to a standard. Index values of 0.7 or higher indicate that items are distributed evenly among assessed standards.

**These cutoffs were offered as example minimums, based on a hypothetical test context. Although appropriate criteria and cutoffs should be considered for each specific test context, stakeholders tend to use these example minimum thresholds in practice. If test blueprints specify intended depth, breadth and balance, then these specifications can help define appropriate thresholds for each alignment criterion.*

For purposes of this study, the absolute minimum threshold was set for each of these four criteria. Panelists additionally evaluated the extent to which the PoP design elements of equity, authenticity and student agency were evident in the content of the assessment. In addition to the evaluation of the assessment tasks, a high-level framework analysis was conducted to determine the extent to which the overall structure of the performance assessment had the capacity to meet alignment expectations as represented in current federal regulations.

In the first year of pilot administration, each PoP included three tasks. Complete assessments (sets of three tasks) were available for mathematics grade three and for English language arts grades six, seven and eight. Content analyses were conducted remotely by three-person expert-educator subject area panels. Reviewers had experience with both classroom and state-level performance assessment development, review and/or implementation as well as with content alignment analysis. Panelists accessed all assessment materials via Google Drive and conducted meetings via Zoom video conferencing. Data were recorded in an online data collection system connected to a server at the Wisconsin Center for Education Research.

For an alignment analysis of a standardized assessment, the unit of analysis is typically the test item. For a performance assessment, selecting the appropriate unit of analysis for evaluation may take some consideration to determine the appropriate grain size as well as which components, if any, to exclude. As a curriculum-embedded assessment, some PoP components served to spark student interest or support project planning but were not intended for evaluation or grading. For purposes of conducting an alignment analysis that would parallel federal expectations, only the assessment prompts used to elicit work that contributed to the student's overall determination of performance (i.e., work that was scored) were identified as the units of analysis.

Rubrics can help identify appropriate units of analysis for content alignment analyses of performance tasks with assessment targets. The pilot administration PoP used one-point rubrics that provided a high-level link between the performance tasks and each standard assessed. Detailed, analytic rubrics are important to include in an alignment analysis if used for scoring. These rubrics convey the relative weighting of tasks and make clear which components of work contribute to a student's score. Further, alignment evidence should demonstrate not only that the tasks provide students the opportunity to demonstrate expectations of the standards but also that the determination of proficient performance, as represented within a rubric, links back to both the task and the intended measurement targets. The inferences about achievement that are made based on student work are commonly made visible through rubrics. Thus, an aligned system requires consistency between and among the academic standards, the opportunities students are provided to demonstrate mastery (i.e., assessment tasks) and the rubrics used to evaluate student work.

Federal peer review is evidence-based and retrospective. In the case of alignment analysis for purposes of submission of evidence for federal peer review, findings are based on the assessment that was administered to students in the previous school year. If an alignment gap is known or identified, that alignment gap is considered an evidence-based gap for the particular year of administration, even if plans are in place to resolve the weakness in future years. Similarly, if a state has other known weaknesses in the technical quality of an assessment, any plans to address these weaknesses do not alter the findings that apply to the already administered assessment. In the context of federal peer review, an alignment analysis is conducted with an operational assessment that is expected to have gone through thorough development and review processes and has already been used to make consequential decisions about student proficiency. For this exploratory work, MCIEA's PoP was used as a case study, and the analysis was not conducted for purposes of a state's submission to peer review. Instead, the purpose of the alignment analysis was to consider the compatibility of a curriculum-embedded performance assessment with federal requirements, using the PoP pilot as an example for purposes of illustration. As such, panelists were instructed to allow for slight adjustments or corrections to the assessment components when conducting their evaluation. If panelists found that a task had clear potential to elicit what students know and can do as relates to a standard, but the task needed some revisions, the task was still coded affirmatively. Panelists recorded these allowances in their item-level notes.

Overall, analysis of the data resulting from the alignment evaluation yielded the typical information necessary for submission to assessment peer review about the degree to which the PoP assessments were aligned with the expectations outlined in the corresponding standards, with consideration of the four commonly used content-focused criteria of Categorical Concurrence, Depth of Knowledge (DOK) Consistency, Range of Knowledge Correspondence and Balance of Representation. Analyses also yielded information beyond the requirements of existing federal regulations and specific to the performance assessment context, with a focus on the research questions guiding this work.

Alignment Findings

Framework Analysis: Overall Capacity for Alignment

In the early-stage pilot form, much work was still in progress, and limited assessment framework documentation was available. One purpose of the PoP assessment program is to provide a determination of student proficiency according to four proficiency levels, like the statewide MCAS. Additional clear and thorough documentation of the purposes and uses of the assessment, along with test rationale (e.g., a theory of action), would provide helpful grounding against which program evaluation could be made.

Overarching valued test features of relevancy, interdisciplinary connections, student choice and voice, equity and learning inside and outside the classroom were identified within available assessment framework materials. These features directly reflect goals that are also detailed within the state's Curriculum Frameworks (academic standards). The PoP program rationale could be strengthened by explicitly identifying the alignment relationship between these test features and

the corresponding goals that are represented in the state academic standards which were developed with the input of stakeholders from across the state and adopted by the Massachusetts Board of Elementary and Secondary Education. The PoP framework identifies a core role of purposeful professional learning groups. Again, the program rationale could be strengthened by explicitly linking this feature of the PoP system to the overall goals of the assessment's theory of action, and consistent with the overarching standards reform theory of action, which includes promoting improvements in teaching and learning. These characteristics of the PoP assessment framework design are strong standards-aligned assets.

Capacity for Addressing the Depth and Breadth of the Standards

A premise of our standards-based education system is that we can specify, through the language of the standards, the types of cognitive engagement with academic content that we expect students to have the opportunity to learn. Then, by interpreting the language of the standards, educators and other stakeholders can all "unlock" the same information about the intended complexity of engagement expected. In practice, however, perceptions of complexity are based on conceptual constructs that may vary for many reasons. Even expert educators bring different ideas to the table. To operationalize the intended cognitive engagement as expressed in the standards, educators, content developers and other stakeholders benefit from the use of practical frameworks and tools. The PoP framework expects teachers to attend to the idea of "depth" or "cognitive complexity" by using the lens of the DOK framework to develop tasks that require complex engagement with academic content. DOK is a powerful tool that can be used to promote coherence by helping stakeholders evaluate and communicate about qualitatively different types of expectations and tasks. Just like other common categorization systems, DOK is used to help clarify and communicate about an elaborate construct.

Although a simplification of the more elaborate construct, if a categorization framework is grounded in meaningful distinctions, then the use of the framework can help to enrich understandings of the fuller construct. Importantly, DOK helps educators differentiate task difficulty from task complexity. How one envisions academic expectations, learning and performance as they relate to difficulty and complexity influences how progress and achievement are measured as well as how teachers use assessment results. While all students are to be provided access to the full complexity of cognitive engagement as expressed in the standards, aspects of these expectations may be more difficult for some individuals than for others. Expectations themselves vary in difficulty. Attention to complexity as distinct from difficulty can help ensure all students are provided with access to learning and assessment opportunities consistent with the complexity of engagement specified in the academic standards for all students.

The PoP framework's attention to DOK in the task planning and development process sets up a structure with strong capacity for alignment. Currently, the PoP system emphasizes only high complexity tasks. While the grade-level standards include qualitatively different types of expectations, including both higher and lower complexity expectations at all grade levels, a focus on higher complexity tasks may still be appropriate, as high complexity tasks can inspire student interest and motivation. While these shifts in task complexity may influence score comparability,

comparability at the level required for proficiency determinations can still be supported, even if exact score equivalence is not the goal, given the system's broader instructional aims. Further, engaging students with high complexity tasks can improve outcomes, including as relates to basic skills (Christopherson & Webb, 2024; Darling-Hammond et al., 2020). Continued attention to building a calibrated understanding of the relationship between the cognitive demands of the standards, those of the tasks and opportunities by which students demonstrate mastery of the standards could allow for greater intentionality in task design and help promote greater coherence.

Within the available PoP framework materials, the high-level suggestions for the selection or development of tasks provide a structure that could reasonably meet expectations for breadth consistent with current federal expectations. For example, teachers are advised to identify standards for assessment that are drawn from across the discipline's domains, or conceptual categories (i.e., intentionally selecting across the breadth of the standards). PoP guidance refers to this set of standards as "power standards." While the intent is to focus efforts on high-leverage content, it is possible that the emphasis on a selected set of "power standards" could unintentionally limit the curriculum or overemphasize a small set of standards. Alternatively, teachers could be advised to consider the way(s) that the standards are grouped into larger categories and to identify the way(s) in which the expectations of the standards intertwine and interweave. The English language arts standards, for example, emphasize an integrated model of literacy and direct educators to examples of "how standards may be combined in effective instruction." The mathematics standards note that expectations are organized into "critical focus areas" that can be helpful for designing curricula. Eight "Guiding Principles" and eight "Standards for Mathematical Practices" emphasize the importance of an integrated and cohesive approach to mathematics education. Emphasizing groupings or collections of standards that include knowledge and skills that work together to build college- and career-ready students could help promote the development of rich, authentic tasks, in line with program goals.

Overall, program guidance suggests the development of three to five tasks per grade and subject area. If each of these tasks requires knowledge and skills drawn from multiple grade-level standards and provides opportunities for students to demonstrate the same complexities of engagement as represented in the standards, then the general structure can yield assessments aligned with the corresponding sets of grade-level standards. Refinements to the descriptions for the program-defined "elements of quality" could additionally help promote intentionality and greater coherence.

In general, test framework materials suggest capacity for alignment, but additional detail and rationale throughout would strengthen this capacity. This finding should not be interpreted as a weakness of the program but rather as an indication of the powerful potential for such an assessment program to promote the growth and development of school faculty through collaborative practice and reflection. For assessments to be a tool for the transformation of teaching and learning, as intended by standards-based reform, our findings emphasize the corresponding need to allow sufficient time for such processes to develop and take shape.

Portfolios of Performance Task Sets: Cutoffs for Alignment Criteria

For each grade and subject area PoP, alignment findings were reported by domain and by criterion (Appendix A). In addition, an overall determination of alignment is typically provided for each assessment. A test form (or other type of assessment event) may not meet specific alignment criteria for one or more domains but still be acceptably aligned overall based on the cutoffs described below. Overall alignment is typically reported based on the following categories:

- **Fully meets expectations:** A summative assessment test form *fully meets* expectations for alignment with corresponding standards if no changes are needed to meet agreed-upon minimum cutoffs for all alignment criteria.
- **Acceptably meets expectations:** A summative assessment test form *acceptably meets* expectations for alignment with corresponding standards if it needs up to 10% of included items revised or replaced to meet all cutoffs. For example, a test form that has 50 items and that needs between one and five items revised or replaced to meet all cutoffs would be considered to acceptably meet alignment expectations.
- **Needs slight adjustments:** A summative assessment test form *needs slight adjustments* if it needs between 10% and 20% of included items revised or replaced to meet all cutoffs.
- **Needs major adjustments:** A summative assessment test form *needs major adjustments* if it needs over 20% of included items revised or replaced to meet all cutoffs.

These categories have been commonly used in the context of submission to federal assessment peer review, and the widely accepted decision rule was grounded in the context of a typical multiple-choice test form of around 50 items that were generally equally weighted. These same proportional thresholds can be applied to a variety of test designs and can take weighting into account if applicable.

Overall alignment is not reported for the PoP assessments because specific scoring information, such as any relative weighting of task components, was not available at the time of data collection. In general, based on the content analyses conducted, all assessments would need some adjustments to meet all four alignment criteria used in this analysis. However, all assessments could meet all alignment criteria with these adjustments.

Portfolios of Performance-Specific Criteria: Equity, Authenticity and Agency

Across grades, most tasks reviewed either met or had the potential to meet the assessment's intended elements of quality: equity, authenticity and agency as defined by the PoP program. Compared with the other sets of tasks, panelists found the grade eight tasks to most strongly embody the elements of equity, authenticity and agency. Panelists commented that in some cases, the links between the tasks and the design elements of equity, authenticity and agency met the letter of the program's definition but not the spirit.

For example, students may have been provided the opportunity to make a choice related to their work, but panelists questioned whether the choice provided met the intention of the criterion of student agency as defined by the test program. Findings suggest that additional work is needed to help teachers calibrate their understanding of these characteristics and develop strategies to incorporate them into tasks.

Again, as the assessment was in early stages of development, this finding should not be interpreted as a program weakness; rather, it underscores the potential for the assessment to positively influence teaching and learning as well as the need to allow sufficient time for program development.

Beyond these criteria explicitly defined within the PoP assessment framework, reviewers noted several additional observations related to the alignment of standards and the PoP assessment. For example, they noted that the English language arts performance tasks reflected the "integrated model of literacy" structure of knowledge as outlined in the state's curriculum framework (standards), in which standards from multiple domains would be interwoven rather than addressed in isolation. They also noted the cross-curricular links within many of the tasks, consistent with the interdisciplinary goals identified in the state standards. The performance tasks also provided students with opportunities that closely reflected the intended pedagogical implications of the standards, such as the strategies of formal analysis and comparative analysis for critical reading, opportunities for longer writing compositions that require extended, reflective and iterative work that could not be completed in one sitting and collaborative work with peers.

Comparability Methods

The IADA provides an opportunity for states to pilot an alternative measure of student achievement in a subset of districts with the intent that the alternative measure will ultimately be used in the state's Title I accountability system. While states must have a plan to scale the innovative assessment system statewide, the IADA provides significant flexibility for states to pilot and improve its innovative assessment of student achievement before engaging in large-scale implementation of the design. During the pilot period, states must demonstrate that the innovative assessment system produces annual determinations of student achievement that can be reasonably compared to the existing summative assessment. The comparability requirement ensures that schools participating in either innovative or legacy assessment systems can be evaluated alongside one another fairly within the same state accountability system. For innovative assessments of student achievement to be deemed comparable, three levels of score comparability must hold, as shown in Figure 5.

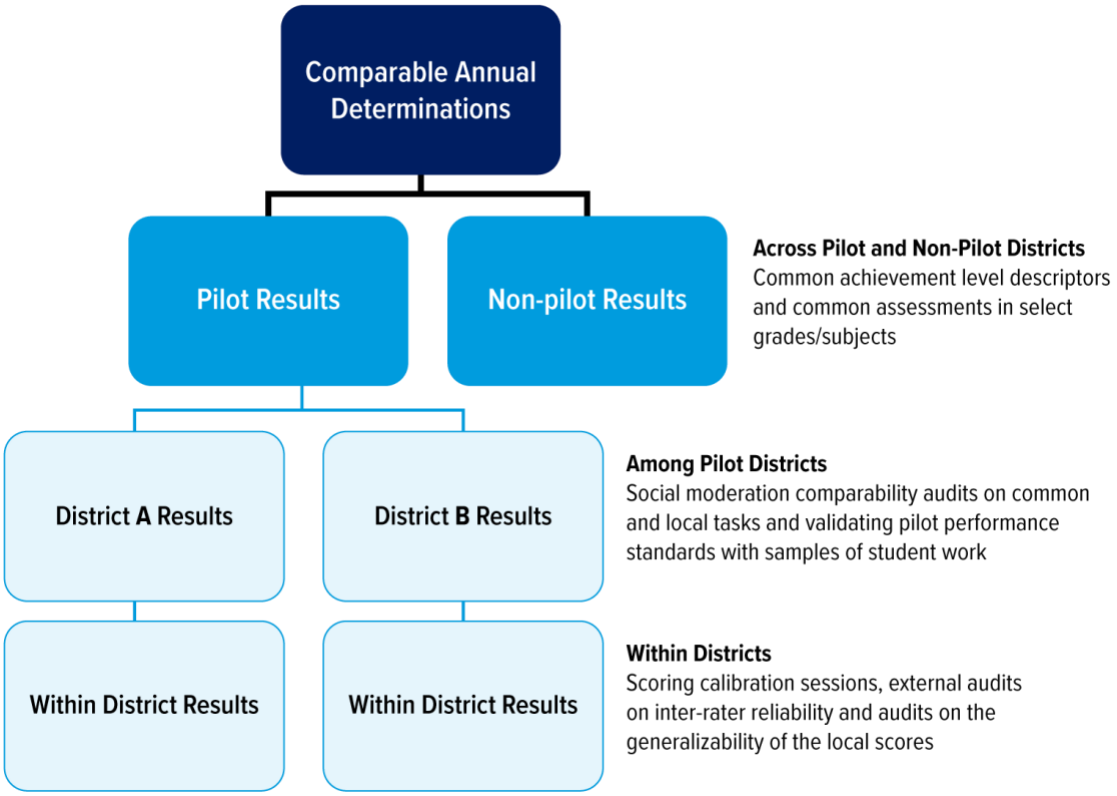

***Figure 5**. Establishing an evidence base for comparable annual determinations (Lyons et al., 2016)*

There were three primary sources of data required for the planned data analyses related to score comparability:

1. **MCIEA educator-rated portfolio scores for all participating students**
   These data were shared with the Lyons Assessment Consulting research team in July 2024. Scores were de-identified to protect student privacy but maintained unique Research IDs for matching students across data sources. These scores, assigned by a student's teacher of record, are referred to in this report as "educator scores."
2. **Double-blind portfolios scored by other educators**
   To examine score comparability among pilot districts, we leveraged the data from the double-blind scoring process that MCIEA has in place. This process engages MCIEA educators in blind scoring student portfolios from districts other than their own. In this report, these scores are referred to as "double-blind scores." The scores are captured in Pearson's Online Scoring Control and Reporting (OSCAR) scoring system, which has been developed to securely maintain student portfolios and facilitate double-blind scoring for the MCIEA project. These data were shared with the Lyons Assessment Consulting research team in July 2024.
3. **Spring 2024 MCAS scores for all participating students**
   These data were shared with the Lyons Assessment Consulting research team in September and October 2024. Scores were de-identified to protect student privacy but maintained unique Research IDs for matching students across data sources.

To evaluate the degree of compatibility of the MCIEA PoP assessment system with federal IADA requirements related to score comparability, we conducted two analyses:

1. Evaluating inter-rater agreement rates across MCIEA educator proficiency determinations and MCIEA double-blind proficiency determinations by district, grade and subject
2. Evaluating agreement rates in proficiency determinations across the MCIEA Portfolios of Performance and MCAS by district, grade and subject

All portfolios were scored on a scale of one to four, representing four proficiency levels: Not Meeting Expectations, Partially Meeting Expectations, Meeting Expectations and Exceeding Expectations. For our analyses, we calculated exact inter-rater agreement and the correlation (Kendall's τ) between the ratings/scores from each different source. Differences in the comparability of scores among the assessment systems by student group (i.e., gender, race/ethnicity, individualized education program or IEP status, English language learner status) were explored when feasible, given sample sizes and data availability.

Comparability Findings

A total of 554 portfolios were collected from the four partner districts. From those, 168 students had double-blind scores provided with their educator-rated portfolio scores, and 547 could be linked to a corresponding MCAS score. By subject area, 70% of the portfolios represented English language arts, and 30% represented mathematics. Looking across grade bands, 100% of the English language arts scores came from grades six, seven and eight, and 100% of the mathematics scores came from grade three. A more robust breakdown of the sample and findings can be found in Tables 2 and 3.

Educator and Double-Blind Scores

Across all pairs, there was middling agreement between educator and double-blind scores (54%). However, the relationship was still positively (albeit weakly) correlated ($\tau = .39$). There was somewhat higher agreement (61%, $\tau = .52$) for the middle grades' English language arts portfolios compared with the grade three mathematics portfolios (47%, $\tau = .06$). In a similar set of analyses conducted for the New Hampshire PACE assessment, a threshold of 60% was established as a baseline for acceptable agreement (Lyons et al., 2017). Following this guideline, only one of the partner districts had an acceptable agreement (grade six English language arts; 74%), with a strong correlation between educator and double-blind scores ($\tau = .72$). These results are reported in Table 2 with district names masked to protect participant privacy.

Table 2. Comparison of Portfolio and Double-Blind Scores by Subject and Grade

Subject	District	Grade	N or Sample Size	Portfolio	Blind	% Agree	Correlation
ELA	All	6-8	87	2.83	2.60	60.92	0.52**
Math	All	3	81	2.32	2.10	46.91	0.06
Math	A	3	45	2.38	2.00	44.44	-0.06
ELA	B	6	35	2.75	2.67	74.29	0.72**
ELA	B	7	12	2.73	2.58	58.33	0.47
Math	C	3	36	2.27	2.22	50.00	0.22
ELA	D	8	40	3.15	2.55	50.00	0.51**

**$p < .01$

Across racial/ethnic groups, agreement ranged from 51% to 82%, and correlations were significant and positive, ranging from 0.33 to 0.71. Across other demographic variables, we see similar patterns in correlations (all positive, ranging from 0.27 to 0.41) and agreement (52% to 58%); however, no demographic group surpassed the 60% threshold of agreement. In almost every group comparison, the average double-blind score (from educators in another district) was lower than the average educator score (from the student's teacher of record). Female students outperformed male students in both educator scores and double-blind scores, but this difference was far more pronounced when just considering double-blind scores. On the other hand, for all other demographic group comparisons, the double-blind scores had smaller performance gaps than the educator scores.

Educator and MCAS Scores

The overall correlation between educator scores and MCAS scores was significant and positive ($\tau = 0.50$), but the 44% agreement failed to surpass the 60% threshold. Within middle school English language arts and grade three mathematics, both correlations were significant at 0.56 and 0.28, respectively, but the agreement once again fell short of acceptable at 47% and 36%, respectively. The one district/grade that demonstrated sufficient agreement was grade eight English language arts at 64%. These results are reported in Table 3.

Table 3. Comparison of Portfolio and MCAS Scores by Subject and Grade

Subject	District	Grade	N or Sample Size	Portfolio	MCAS	% Agree	Correlation
ELA	All	6-8	388	2.83	2.43	47.42	0.56**
Math	All	3	159	2.25	2.16	35.85	0.28**
Math	A	3	81	2.38	2.26	34.57	0.24**
ELA	B	6	230	2.75	2.33	44.35	0.54**
ELA	B	7	81	2.73	2.09	40.74	0.59**
Math	C	3	78	2.27	2.05	37.18	0.33**
ELA	D	8	77	3.15	3.21	63.64	0.60**

**$p < .01$

Within nearly every subgroup educator scores were higher than MCAS scores. For all racial/ethnic categories, the correlations were significant and positive (0.44 to 0.61), but none demonstrated sufficiently high inter-rater agreement (30% to 51%). A similar set of results was also found within other demographic comparisons (31% to 47% agreement, 0.28 to 0.53 correlation). Students receiving free or reduced-priced lunch, English language learning services or special education services had lower portfolio scores and lower MCAS scores on average than those not receiving services. Notably, the agreement between scores for these groups was also observably lower (7% to 14% lower). Performance gaps associated with each demographic group were smallest when comparing double-blind scores and greatest when comparing MCAS scores. A full table of results broken down by subgroup can be found in Appendix B.

As a post-hoc analysis, we explored the relationships between the double-blind scores and the MCAS scores. In most cases, these relationships were weaker than the correlations between educator-scored portfolios and MCAS scores. Taken together with the other results of our analysis, we believe that the current MCIEA PoP test design does not provide enough information to produce consistently scorable proficiency scores for students. However, this is an early-stage pilot, and revisions to the design may produce higher agreement in future iterations. High agreement rates between portfolio scores and double-blind scores are a necessary condition for establishing the independent reproducibility of scores. Without sufficient agreement at the program level, the low agreement rates and correlations with the MCAS scores are an anticipated finding, as there is likely too much noise in the MCIEA PoP scores to systematically relate to an external measure.

Discussion of Findings

At a high level, the sample of PoP assessments included in this study either met or showed potential to meet existing federal expectations for content alignment. Findings suggest that the MCIEA assessment system is not yet mature enough to meet the technical quality requirements of the law as relates to comparability. Although the PoP assessments did not fully meet federal expectations in pilot form, it is worth noting that mature, resource-intensive commercial testing programs also do not necessarily fully meet federal expectations.

While USED's 2023 request for information suggested that states perceived federal expectations for alignment as a barrier to innovation, findings from this study suggest that existing federal alignment expectations are reasonably attainable for a performance assessment. Authentic problems involve multiple layers of knowledge and skills, drawing from a range of grade-level standards and provide the opportunity for students to engage in a variety of complex ways with academic content. Beyond the capacity to meet expectations for depth and breadth, performance assessments provide an opportunity to address other key alignment criteria that are not represented in current federal expectations. Reviewers noted a variety of observations related to factors such as the structure of knowledge and the pedagogical implications of the assessments that are not evident through the lens of existing federal expectations. For example, the assessments included cross-curricular connections and collaborative work, just as the standards communicate these as valued goals that should be attended to in the classroom environment. In general, the assessment formats and structure were similar to the types of effective classroom practices that the standards were intended to promote.

For score comparability, most PoP assessments reviewed did not meet the minimum criteria. For MCIEA, the following three recommended programmatic changes are likely to significantly improve the program's score comparability (and could also help strengthen alignment):

1. **Increasing the Number of Tasks**: The portfolios reviewed in this study included only three performance tasks within a given grade and subject area. The score comparability evidence suggests that the limited number of tasks does not offer a full enough picture of student achievement for educators to reliably and accurately assess student proficiency.
2. **Strengthening Task Quality**: Given this was a pilot year for the MCIEA assessment system, many of the participating educators indicated it was their first time developing, administering and scoring performance tasks. Designing high-quality performance tasks that bring students into meaningful interaction with the assessment content is a technical skill that can require years of training and practice. As such, we recommend that MCIEA leverage its robust bank of tasks that have been developed and reviewed by trained educators for future years of the assessment system pilot.

3. **Providing Additional Scoring Training and Supports:** Increasing the number of tasks and strengthening task quality will certainly improve educators' ability to accurately and reliably assign a level of proficiency. Additional supports in scoring, such as providing multiple opportunities to practice with the rubric, running cross-district calibration sessions and engaging in consensus scoring, are also likely to improve score comparability.

The MCIEA team incorporated these recommendations along with additional program improvements to strengthen their pilot in the 2025–2026 academic year. This underscores the value and importance of funding research-practice partnerships like this one for the advancement of innovative assessment models.

Despite the limitations of the current MCIEA model for meeting technical quality requirements, it is worth noting that the program has a number of strengths related to the theoretical coherence of the assessment design with the intended deeper learning outcomes. For example, most of the tasks met (or, with some revisions, had the potential to meet) program-defined criteria related to equity, authenticity and student agency. Further, the curriculum-embedded performance assessments provided opportunities for students to engage in tasks that reflected the intended structure of knowledge as well as pedagogical implications of the standards. As the MCIEA model matures and strengthens both its assessment design and factors related to program implementation, our findings suggest a strong capacity to build and support a comprehensive validity argument in support of a performance-based assessment system.

Charting a Path Forward: Policy Recommendations

Recommendation One:
Strengthen Support in the Field for the Development of Performance-based and Other Innovative Assessment Models

The case study found that the Massachusetts Consortium for Innovative Education Assessment's (MCIEA) model had the potential to meet federal requirements, but only with focused iteration and program investment. This finding is consistent with the MCIEA team's expectations of results for their first-year pilot. As such, the program applied findings from this study to develop its professional learning plans and model adjustments for the 2025–2026 school year. This finding exemplifies the need for practitioners, policymakers and researchers to collaborate to create conditions for the successful development and implementation of performance-based assessment models. The following actions could support this recommendation:

Develop a Structure for Sharing Innovative Assessment Models and Lessons Learned Within and Across State Contexts

The MCIEA program team reported limited opportunities for collaboration with the state education agency during the development of their performance assessment model. The state agency is in the process of developing science performance assessments under an Innovative Assessment Demonstration Authority (IADA) authorization. Several districts involved in the case study are participating concurrently in both pilots. Both state and federal policymakers should create programs or learning networks to facilitate documentation and sharing of lessons from the development of innovative assessment models. Such structures provide space to document key assessment design tradeoffs, minimize replication of common design flaws, encourage the adoption of best practices and improve coherence for the field. Throughout our research process, the results indicated the need to bridge content, measurement and policy siloes.

Invest in Innovative Assessment Models

The study findings helped inform several strategies for strengthening the quality of the first-year pilot's assessment model, underscoring the value of research and continuous improvement. The federal Competitive Grants for State Assessments (CGSA) program is a valuable mechanism for funding states to develop and refine new assessment approaches. The past two funding cycles have prioritized states seeking to design performance assessment models. Federal leaders should continue to invest in CGSA, increase the pot of available funds and hold the competition annually to ensure states have regular opportunities to apply. Additionally, federal leaders should continue to prioritize the design of performance assessments that support deeper learning in future competitions. At the state level, policymakers should leverage federal opportunities and invest additional resources in the development of performance assessments. The MCIEA program team noted the significant lift asked of teachers and leaders in participating schools. Sufficient resources are needed to both develop and sustain the implementation of performance assessment models.

Support Research on New Innovative Assessment Models

The findings of the case study indicate the need for additional research as well as the expansion of evidence considered in the evaluation of the technical quality of innovative assessments. In the first sections of this paper, the research team suggests reorienting the federal peer review process around well-structured validity arguments. Recognizing the central importance of validity argumentation in modern measurement theory and practice through peer review would create an environment that could incentivize attention to the design tradeoffs and impacts of current common approaches to statewide summative assessment and invite more piloting and research in testing innovation. Policymakers should continue to expand programs that create ample time for piloting and iteration so that states may integrate research as they refine new models.

Recommendation Two:
Create the Conditions to Improve Performance Assessment Quality Within the Current Federal Framework

Study findings highlight several key areas of improvement for state and federal leaders prior to reimagining the current federal framework. Recommendations include actions for state policymakers working to develop performance assessment models, as well as federal policymakers interested in supporting state innovation through adjustments to the current federal program.

Explore the Relationship Between Content Standards Frameworks, Test Blueprints and Instructional Utility

As defined in §200.105(b)(2)(i), federally required assessments must address the "depth and breadth" of the "challenging State academic content standards, for the grade in which the student is enrolled." At the time of the case study, the research team noted a perceived barrier from MCIEA around addressing these alignment criteria. Similar perceptions were expressed to the United States Department of Education (USED) by other states. The research team noted that this barrier may be more of a perceived barrier than an actual barrier. One reason for this perceived concern may be related to a common misconception that USED requires assessments to address every grade-level standard within a subject area. In contrast, blueprints for statewide assessments generally include only a subset of the grade-level standards and, of those, include assessment items and tasks for an even smaller subset of standards. Thus, some of these concerns may be assuaged by the opportunity to learn more about federal expectations and see examples of what has been considered acceptable in terms of meeting alignment expectations. Particularly for English language arts, statewide assessment blueprints typically exclude standards that are considered unassessable in the context of the standardized, on-demand format. Performance assessments may offer potential to assess a greater breadth and variety of standards, including those expectations that involve collaboration, research, internet access, extended time and multiple modes of expression, which are generally excluded from on-demand assessment contexts.

Standardized, on-demand statewide assessments, including the Massachusetts Comprehensive Assessment System (MCAS), are generally structured with a one-to-one ratio between items and standards. In other words, each item is intended to address knowledge and skills from within a single standard. If this same assessment structure were to be directly translated into a performance assessment, then it may indeed be challenging to address the breadth of the standards in a reasonable time span. However, state standards are typically not intended to be addressed individually in instruction and often include explicit statements about the importance of interweaving the individual expectations to meet the higher-level intent of the overall standards. Further, a state's grade-level academic standards statements are often presented in the context of a broader vision of what it means to be on track for college and career readiness, prepared for the 21st century or equipped with a world-class education. Rather than addressing individual content standards with individual tasks, a core rationale for the use of performance assessments is that they offer opportunities for students to engage more fully with the disciplinary structure of knowledge,

consistent with the intent of academic standards' frameworks. If performance assessments are designed around authentic problems and tasks, they will likely address "bundles" of standards, interweaving knowledge and skills within domains, across domains and even across disciplines. Concerns about alignment with academic standards may be mitigated through opportunities to see and experience exemplary performance assessments.

Additional research is needed to better understand the sources of these reported concerns. Results of this exploration may have implications for the design of standards frameworks or teacher professional development around designing assessments and utilizing the data that come from assessments.

Deepen Task and Item Banks

The Massachusetts Consortium for Innovative Education Assessment Task Bank contained 125 tasks at the time of this writing and is expected to grow. These tasks are educator-created, field-tested and reviewed by MCIEA coaches using a program-specific review checklist. Publicly available task and item banks provide educators with vetted tasks that can be implemented flexibly within a teacher's scope and sequence. To ensure that tasks and items within these banks are of high quality, policymakers should work with researchers and program staff to develop validation processes to ensure consistency with nationally recognized approaches to content development and technical quality. Policymakers should make resources available for teachers to contribute to task development, for content and technical quality reviews and for professional development to support teachers implementing performance tasks sourced from task banks.

Develop High-Quality Scoring Practices

The case study identified scoring practices as an area of growth for the MCIEA program. Teachers utilized a one-point, task-neutral rubric that identified the standards intended to be assessed in each task and then holistically scored the portfolios using MCAS rubrics. In addition to supporting reliability in scoring, detailed, analytic rubrics are generally important to include for alignment analyses of performance assessments. Content alignment should be examined across all links, from standards to tasks to rubrics.

Teachers in the program had limited training on the MCAS rubric used to score the performance assessment portfolios. The development of scoring tools and processes is critical to the development of a high-quality performance assessment model. Policymakers should ensure that design tradeoffs are considered carefully, including utilization of task-neutral or task-specific rubrics, teacher scoring, machine scoring (including investigation of artificial intelligence or AI capabilities) and scoring protocols.

Pilot Performance Assessments in Subjects Without a Federal Testing Requirement

Policymakers should consider developing performance assessment models in subjects or contexts that are not subject to current federal requirements. As defined in §200.5, states must administer assessments in English language arts and mathematics every year in grades three to eight and once in high school, as well as one science assessment in each grade band. A key finding of the study is the importance of time and interaction in the development of a high-quality performance assessment model. This is compounded in assessment pilots in federally required subjects that must demonstrate comparability through double testing of the innovative and traditional assessments. Development of performance assessment models in subjects that are not federally required (for example, in social studies, civics, off-year science, portfolios to meet state graduation requirements, etc.) can create the space for the design iteration required for the development of a new assessment model. States may be well-positioned to extend the model to federally mandated subjects once it achieves sufficient quality and the field demonstrates capacity for effective implementation.

Implement Key Technical Improvements to the Innovative Assessment Demonstration Authority

The Innovative Assessment Demonstration Authority (IADA) is the federal program that provides states with flexibility to pilot new assessment models. Application requirements outlined in §200.105 include requirements for demonstration of comparability between the state's current system and the newly designed innovative system. National technical experts agree that overemphasis on comparability in this way can constrain innovation that aims to design assessments better than those currently implemented (Lyons & Marion, 2016). There are five approaches states may use under current regulations to meet comparability requirements. These requirements all involve, in some fashion, generating comparable results across the state's innovative assessment and the corresponding federally mandated statewide assessment. While the need to have reliable data for making determinations about the performance of students across a state remains, the current approach to comparability constrains the design of the innovative assessment to the design and reporting structures of the state's existing statewide assessment. This limits the ability of states to create better assessments through IADA that capture more authentic, expansive or enhanced information about what students know and can do.

Another critical concern with the current IADA program is the requirement that states be ready to scale in five years. This research shows that model development requires significant time and iteration. We recommend that federal policymakers:

- Consider explicitly allowing for a time-limited focus on comparability of the proficiency category only within the demonstration period, rather than the requirement that comparability be established in each year of the pilot phase. In subsequent years, states should have the option to demonstrate comparability of newly onboarded schools and districts to the new system (Lyons & Marion, 2016).

- Examine the impact of how comparability is defined for IADA purposes. Rather than expecting strict score comparability between a state's existing statewide assessment and an IADA assessment, we urge USED to consider score comparability requirements for IADA assessments with consideration of the important potential differences (between existing statewide and IADA assessments) in the alignment relationships with the state's standards.
- Increase program participation by creating a planning period and permitting states to propose a more realistic timeline for scaling administration that considers the complexity of the assessment design and size of the state's student population.

Recommendation Three:
Reorient the Federal Assessment Paradigm to Enable and Facilitate Assessment of Deeper Learning

To truly encourage performance assessment models at scale, federal policymakers will need to partner with states and the research community to reimagine key aspects of the nation's paradigm for state assessments. The first actionable step listed below recognizes an opportunity to develop coherence among national programs intended to encourage assessment innovation and could be executed in the relatively near term. The remaining actionable steps would require federal policymakers to invest in the development of new assessment models with a deeper learning orientation.

Integrate Federal Programs for Assessment Innovation

At the time of writing this paper, the federal government offered two programs for states seeking to innovate the models of their state assessment system. The CGSA program is a semi-regular competition where states can apply for financial resources to support improvements to their state assessment systems. As noted previously, the Innovative Assessment Demonstration Authority (IADA) program allows states to innovate federally required state assessments but does not provide associated resources. While these programs share similar goals and are often utilized by the same states, they are not formally connected. Federal policymakers should consider integrating these programs to provide greater coherence for participating states.

Study Sampling Models to Reduce Testing Burden

In debriefing study findings, the MCIEA program team noted the significant capacity necessary to maintain a system of deeper learning assessments at the frequency currently required by §200.5. The combination of public concern over the amount of testing under the current model and the implications of administering that same frequency of testing with deeper assessments raised important questions about feasibility. Federal policymakers may wish to consider reimagining the requirements to assess every learner every year. Recent analysis has indicated the feasibility of utilizing sampling models to reduce testing burden, and policymakers should engage the research community to study this question more deeply as they explore conditions for successful performance assessment models (Marion et al., 2024). Adjustments to the federal assessment paradigm that would support curriculum-embedded assessment may also help with this issue.

Develop a Dynamic Methodology to Evaluate Assessment Technical Quality in the Context of a Well-structured Validity Argument

This case study provided our study team the opportunity to examine metrics emphasized by nationally accepted methodologies for evaluating assessment quality. The study utilized the four Webb alignment criteria that are most used for this context and the *Standards for Psychological and Educational Testing* (2014) as the basis for the alignment and score comparability studies, respectively. While both frameworks articulate a broad range of indicators of technical quality for assessment systems, federal peer review focuses on a narrow

set of those indicators. For example, Webb articulated 12 criteria in the original 1997 framework for evaluation of alignment, but current practice for large-scale summative assessments commonly focuses on just four of those criteria.

Key criteria that could be used to evaluate the transformational use of assessment to improve teaching and learning, including alignment criteria related to pedagogical implications and, more broadly, validity evidence based on test consequences, are not currently included in federal peer review expectations for large-scale summative assessments despite their inclusion in these national frameworks for assessment quality. The study team noted the need for a methodology that widens the scope of evidence that is valued regarding the technical quality of assessments and recognizes the importance of well-structured validity arguments within modern measurement theory and practice. Federal policymakers should convene content, learning sciences and measurement experts to develop an adaptable methodology that values a strong theoretical grounding for technical quality, creates space for states to develop an approach consistent with their theory of action and presents evidence within a comprehensive validity argument with defensible logic behind tradeoffs in design decisions. This change would need to be reflected in the peer review guidance and review processes.

Conclusion

This case study explored one model for performance assessment and examined the compatibility of this innovative assessment with current federal requirements. The research also provided the study team with the opportunity to explore how federal requirements could be reimagined to honor and encourage more meaningful assessment systems. By prioritizing validity argumentation in federal peer review, we could reorient the evaluation of technical quality to emphasize the role of a clear test rationale (such as a theory of action), the importance of theoretical coherence, the value of expanded evidence of alignment and score comparability, the need to monitor integrity of implementation and to evaluate systemic impacts. Our results provided feedback to the Massachusetts Consortium for Innovative Education Assessment team to improve their program as well as key insights for how practitioners, policymakers and researchers can partner to evolve our assessment systems toward deeper learning.

Acknowledgments

This report was funded as part of a collection of research on innovative assessment approaches and alternative accountability models coordinated under The K12 Research for Equity Hub. The Hub is managed by EduDream and funded by Gates Foundation and Walton Family Foundation. No personnel from Gates Foundation or Walton Family Foundation participated in the creation of Hub research. The findings and conclusions contained in this report are those of the authors and do not necessarily reflect the positions and/or policies of Gates Foundation or Walton Family Foundation. Thank you to the EduDream team, the EduDream Cohort 2 participants and the EduDream Research Advisory group for their guidance and feedback throughout this research.

Lyons Assessment Consulting is a leader in supporting innovation in educational assessment and accountability. We offer a full range of custom consulting solutions. Our services include supporting the design of innovative assessment and accountability systems, leading research and producing white papers, and serving as technical advisors to large and complex projects. Our clients range from forward-thinking non-profit organizations and government education agencies to testing companies and influential policy think tanks who partner with the team at Lyons Assessment Consulting to create meaningful improvements in educational assessment and accountability. We work closely with our clients to transform traditional systems to better serve all students.

The Wisconsin Center for Education Products and Services (WCEPS) is a non-profit organization affiliated with the University of Wisconsin-Madison. WCEPS brings research-driven content analyses, program evaluation, educational tools, and transformative professional learning experiences to educators across the country and around the globe. Our programs partner with educational content and assessment developers, schools, districts, state departments of education, and other educational entities to inform, inspire, and ultimately, to co-create positive change that makes a lasting difference in the lives of educators and students. webbalign.org

KnowledgeWorks is a national nonprofit organization advancing a future of learning that ensures each student graduates ready for what's next. For more than 25 years, we've been partnering with states, communities and leaders across the country to imagine, build and sustain vibrant learning communities. Through evidence-based practices and a commitment to equitable outcomes, we're creating the future of learning, together. KnowledgeWorks.org

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (Eds.). (2014). *Standards for Educational and Psychological Testing*.

Applications for New Awards; Competitive Grants for State Assessments Program, 89 FR 16750 (March 8, 2024). https://www.federalregister.gov/documents/2024/03/08/2024-04972/applications-for-new-awards-competitive-grants-for-state-assessments-program

Bennett, R. E. (2010). Cognitively Based Assessment of, for, and as Learning (CBAL): A Preliminary Theory of Action for Summative and Formative Assessment. *Measurement: Interdisciplinary Research and Perspectives, 8*(2–3), 70–91. https://doi.org/10.1080/15366367.2010.508686

Christopherson, S., & Webb, N. (2024). *Complexity and Difficulty in a Coherent Standards-based Education System*. Wisconsin Center for Education Products and Services. https://www.webbalign.org/difficulty-and-complexity

Darling-Hammond, L. (2017). *Developing and Measuring Higher Order Skills: Models for State Performance Assessment Systems.* Learning Policy Institute and Council of Chief State School Officers. https://learningpolicyinstitute.org/product/models-state-performance-assessment-systems-report

Diaz-Bilello, E., & Pierre-Louis, M. (2021). *Documenting the Implementation and Uses of Performance-based Assessments to Evaluate Postsecondary and Workforce Readiness*. University of Colorado Boulder Center for Assessment, Design, Research and Evaluation. https://www.cde.state.co.us/postsecondary/pacasestudypaper2

Darling-Hammond, L., Adamson, F. (2010). *Beyond Basic Skills: The Role of Performance Assessment in Achieving 21st Century Standards of Learning*. Stanford Center for Opportunity Policy in Education. https://edpolicy.stanford.edu/sites/default/files/publications/beyond-basic-skills-role-performance-assessment-achieving-21st-century-standards-learning_4.pdf

Englert, K., & Shultz, P. K. (2025). Reconsidering Theories of Action: Using Cultural and Community Validity to Transform Assessment Development, Implementation, and Use. In C. M. Evans & C. S. Taylor (Eds.), Culturally Responsive Assessment in Classrooms and Large-Scale Contexts Theory, Research, and Practice (pp. 156–176). Routledge. https://doi.org/10.4324/9781003392217-12

Fulmer, G. W. (2011). Estimating Critical Values for Strength of Alignment Among Curriculum, Assessments, and Instruction. *Journal of Educational and Behavioral Statistics*, 36(3), 381–402. https://doi.org/10.3102/1076998610381397

Fulmer, G. W., Tanas, J., & Weiss, K. A. (2018). The Challenges of Alignment for the Next Generation Science Standards. *Journal of Research in Science Teaching, 55*(7). 1076–1100. https://doi.org/10.1002/tea.21481

Ihlenfeldt, S. D., Student, S., Lyons, S., Dadey, N., Forte, E., & Winter, P. (2024). *Enhancing Peer Review: Supporting Innovation in State Assessment Systems.* Lyons Assessment Consulting. https://lyonsassessmentconsulting.com/resource/recommendations-for-supporting-innovative-state-assessment-systems/

Kane, M. T. (1992). An Argument-based Approach to Validity. *Psychological Bulletin*, *112*(3), 527–535. https://doi.org/10.1037/0033-2909.112.3.527

Kane, M. (2004). Certification Testing as an Illustration of Argument-based Validation. *Measurement, 2*(3), 135–170. https://doi.org/10.1207/s15366359mea0203_1

Lane, S. (2014). Evidence of Validity Based on the Consequences of Test Use. *Psicothema, 26*(1), 127–135. https://doi.org/10.7334/psicothema2013.258

LeMahieu, P. (2011, October 11). *What We Need in Education is More Integrity (and Less Fidelity) of Implementation.* Carnegie Foundation for the Advancement of Teaching. https://www.carnegiefoundation.org/blog/what-we-need-in-education-is-more-integrity-and-less-fidelity-of-implementation/

Lyons, S., Evans, C., Marion, S., & Thompson, J. (2027). *New Hampshire Performance Assessment of Competency Education (PACE): Technical Manual*. Retrieved from: https://www.ed.gov/sites/ed/files/policy/elsec/guid/stateletters/nhpacetechmanual72017.pdf

Lyons, S., Marion, S. F., Pace, L., & Williams, M. (2016). *Addressing Accountability Issues Including Comparability in the Design and Implementation of an Innovative Assessment and Accountability System*. https://knowledgeworks.org and https://nciea.org.

Lyons, S. & Marion, S. F. (2016). *Comparability Options for States Applying for the Innovative Assessment and Accountability Demonstration Authority: Comments Submitted to the United States Department of Education Regarding Proposed ESSA Regulations*. Retrieved from https://nciea.org.

Marion, S. (2010). *Developing a Theory of Action: A Foundation of the NIA Response.* Center for Assessment. https://www.nciea.org/wp-content/uploads/2021/11/Theory-of-Action_041610_2.pdf

Marion, S., & Lorié, W. (2024, July 19). *Can We Reduce Testing in K-12 Schools?* Center for Assessment. https://www.nciea.org/blog/can-we-reduce-testing-in-k-12-schools/

Markle, R. (2024). A Call to Action: Integrating Theories of Action as a Modern Component of Validity. *Applied Measurement in Education.* https://doi.org/10.1080/08957347.2024.2445838

Messick, S. (1994). The Interplay of Evidence and Consequences in the Validation of Performance Assessments. *Educational Researcher, 23*(2), 13–23. https://doi.org/10.3102/0013189X023002013

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (2003). On the Structure of Educational Assessments. *Measurement, 1*, 3–62. https://doi.org/10.1207/S15366359MEA0101_02

Montana Alternate Student Testing Task Force & Arntzen, E. (2022). *Assessment Design and Implementation Considerations for the Montana Alternate Student Testing (MAST) Pilot Program*. Montana Office of Public Instruction. https://www.nciea.org/wp-content/uploads/2022/08/MAST-Report-R4.pdf

National Academies of Sciences, Engineering and Medicine, Division of Behavioral, Social Sciences, Board on Science Education, Board on Behavioral, … & Practice of Learning. (2018). *How People Learn II: Learners, Contexts, and Cultures*. National Academies Press.

Richardson, J. (2017). *PDK Poll.* Phi Delta Kappan International, 23–24.

Pape, B. (2018). *Learner Variability is the Rule, Not the Exception*. Washington, DC: Digital Promise Global. https://digitalpromise.org/wp-content/uploads/2018/06/Learner-Variability-Is-The-Rule.pdf

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing What Students Know: The Science and Design of Educational Assessment*. National Academy Press.

Polikoff, M. (2020). The Present and Future of Alignment. *Educational Measurement: Issues and Practice, 39*(2), 18–20. https://doi.org/10.1111/emip.12333

Porter, A. C. (2002). Measuring the Content of Instruction: Uses in Research and Practice. *Educational Researcher, 31*(7), 3–14. https://doi.org/10.3102/0013189X031007003

The Education Trust. (2023). *The Education Trust's Comment Regarding the Innovative Assessment Demonstration Authority*. Available as of December 14, 2023, at https://edtrust.org/press-release/the-education-trusts-comment-regarding-the-innovative-assessment-demonstration-authority/

Shavelson, R. J., Baxter, G. P., and Gao, X. (1993). Sampling Variability of Performance Assessments. *Journal of Educational Measurement 30*(3), 215–232. https://doi.org/10.1111/j.1745-3984.1993.tb00424.x

Student, S. R., Lyons, S., & French, D. (2023). *Performance assessment: A Vehicle for Improving the Utility and Validity of Local and State Assessment Systems*. Education Commonwealth Project. https://static1.squarespace.com/static/64c7dc6e42025770470ca7b1/t/652fd9208c5384629955a590/1697634592612/Performance+assessment+-+A+vehicle+for+improving+the+utility+and+validity+of+local+and+state+assessment+systems.pdf

Subkoviak, M. J. (1988). *A Practitioner's Guide to Computation and Interpretation of Reliability Indices for Mastery Tests. Journal of Educational Measurement, 25(1)*, 47–55. https://doi.org/10.1111/j.1745-3984.1988.tb00290.x

Toulmin, S. (1958). *The Uses of Argument*. Cambridge, UK: Cambridge University Press.

Traynor, A., & Christopherson, S. C. (2024). Using Content Relevance and Representativeness Indices in Instrument Revision. *Applied Measurement in Education*, *37*(2), 132–147. https://doi.org/10.1080/08957347.2024.2347518

Troppe, P., Osowski, M., Wolfson, M., Ristow, L., Lomax, E., Thacker, A., & Schultz, S. (2023). *Evaluating the federal Innovative Assessment Demonstration Authority: Early Implementation and Progress of State Efforts to Develop New Statewide Academic Assessments* (NCEE 2023-004). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. http://ies.ed.gov/ncee

Webb, N. L. (1997). *Criteria for Alignment of Expectations and Assessments in Mathematics and Mathematics Education* (NISE Research Monograph No. 6). Council of Chief State School Officers. Madison: University of Wisconsin, Wisconsin Center for Education Research.

Webb, N. L. (1999). *Alignment of Science and Mathematics Standards and Assessments in Four States* (NISE Research Monograph No. 18). Council of Chief State School Officers; Madison: University of Wisconsin, Wisconsin Center for Education Research.

U.S. Department of Education. (2023). Request for Information Regarding the Innovative Assessment Demonstration Authority, 88 FR 19286. https://www.federalregister.gov/documents/2023/03/31/2023-06697/request-for-information-regarding-the-innovative-assessment-demonstration-authority

Appendix A
Summary of Content Alignment Analysis Data

MCIEA – Portfolios of Performance: Grade 3 Mathematics

The early-stage grade 3 mathematics assessment was focused on a small set of mathematics standards. The task design templates showed that two of the three tasks were each focused on a single mathematics standard and that the third task was focused on two mathematics standards. The narrow focus was likely an artifact of the early stages of development. Reviewers identified several additional standards that were addressed by student work but that were not represented within the task design templates.

No aspect of any task was found to address expectations within the Operations and Algebraic Thinking reporting category. The addition of one more scorable task component that targets a currently unassessed expectation within Number and Operations in Base Ten and Fractions at the corresponding level of complexity would allow for that reporting category to meet minimum thresholds for all four alignment criteria. All alignment criteria were met or weakly met for the Measurement and Data and Geometry reporting category.

Overall, each grade 3 task was designed to focus on just one or two expectations. Another approach for performance assessment is to design a task based on a larger problem context that requires knowledge and skills drawn from multiple standards (and even domains). Some subcomponents of the task may be focused on specific expectations, while others may interweave aspects of multiple expectations.

Table A.1 Summary of Attainment of Acceptable Alignment Level on Four Criteria as Rated by Three Reviewers, MCIEA Grade 3 Math PoP; Number of Assessment Items – 13

Grade 3 Math Reporting Categories	Alignment Statistics				Alignment Findings			
	CC*	DOK %	Range	Balance	CC	DOK	Range	Balance
Operations and Algebraic Thinking (3.OA)	0.0	--	--	--	NO	NT**	NT	NT
Number and Operations in Base Ten and Fractions (3.NBT.NF)	5.0	66%	33%	0.50	NO	YES	NO	NO
Measurement & Data and Geometry (3.MD.G)	16.7	82%	50%	0.67	YES	YES	YES	WEAK

*Number of scored student interactions

** NT = Not tested; no scored student interactions were found to correspond to this domain

MCIEA – Portfolios of Performance: Grades 6–8 ELA

The PoP assessments for grades six to eight focused on reading and writing expectations. Across grades, the assessments met or weakly met all four alignment criteria for the Reading and Writing reporting categories. No scorable task components addressed expectations within the Language reporting category. Reviewers noted that scorable components could readily be added to the existing tasks to address Language expectations.

Table A.2 Summary of Attainment of Acceptable Alignment Level on Four Criteria as Rated by Three Reviewers, MCIEA Grade Six ELA PoP; Number of Assessment Items – 15

Grade 6 ELA Reporting Categories	Alignment Statistics				Alignment Findings			
	CC*	DOK %	Range	Balance	CC	DOK	Range	Balance
Reading (6.R)	18.3	98%	60%	0.85	YES	YES	YES	YES
Writing (6.W)	14.7	88%	57%	0.66	YES	YES	YES	WEAK
Language (6.L)	1.0	100%	17%	1.00	NO	NA**	NA	NA

*Number of scored student interactions

** NA = Not applicable; only one scored student interaction was found to correspond to this domain

Table A.3 Summary of Attainment of Acceptable Alignment Level on Four Criteria as Rated by Three Reviewers, MCIEA Grade Seven ELA PoP; Number of Assessment Items – 15

Grade 7 ELA Reporting Categories	Alignment Statistics				Alignment Findings			
	CC*	DOK %	Range	Balance	CC	DOK	Range	Balance
Reading (7.R)	20.7	62%	63%	0.75	YES	YES	YES	YES
Writing (7.W)	16.3	68%	66%	0.66	YES	YES	YES	WEAK
Language (7.L)	0	--	--	--	NO	NA**	NA	NA

*Number of scored student interactions

** NT = Not tested; no scored student interactions were found to correspond to this domain

Table A.4 Summary of Attainment of Acceptable Alignment Level on Four Criteria as Rated by Three Reviewers, MCIEA Grade Eight ELA PoP; Number of Assessment Items – 21

Grade 8 ELA Reporting Categories	Alignment Statistics				Alignment Findings			
	CC*	DOK %	Range	Balance	CC	DOK	Range	Balance
Reading (8.R)	17.7	51%	85%	0.77	YES	YES	YES	YES
Writing (8.W)	32.0	61%	77%	0.70	YES	YES	YES	YES
Language (8.L)	3.0	61%	30%	0.89	NO	YES	NO	YES

*Number of scored student interactions

** NT = Not tested; no scored student interactions were found to correspond to this domain

Study results for each grade and subject area as summarized in the tables above pertain only to the issue of content alignment between the assessment targets (standards) included in the analysis and the set of PoP tasks that were analyzed. Note that an alignment analysis of this nature does not serve as external verification of the general quality of the standards or assessments.

The full set of study results, including for the PoP-defined criteria of equity, authenticity and agency, as well as all raw data and reviewer comments, were provided to MCIEA. For information on the standards included in the analysis, details of the methodologies employed and details of data collected please contact webbalign@wceps.org.

Descriptions of Each Alignment Criterion Used in this Analysis

Descriptions of the alignment criteria of Categorical Concurrence, Depth of Knowledge Consistency, Range of Knowledge Correspondence and Balance of Representation are provided below. Within these descriptions, the term ***item*** refers to scored student interactions. The word ***domain*** is used to describe the main conceptual content categories as represented within the academic content standards. The term ***standards*** may be used as an umbrella term to refer to expectations in general.

Categorical Concurrence

Aligned academic standards and assessments address the same conceptual or content categories (i.e., domains). The Categorical Concurrence criterion provides a very general indication of alignment if both documents incorporate the same content. *The criterion of Categorical Concurrence between academic standards and assessments is met if the same or consistent categories of content appear in both documents.* This criterion was judged by determining whether the assessment included items measuring content, as explicated in the standards, from each domain. Commonly, an alignment analysis assumes that an assessment must have at least six items (or points for polytomous items) for measuring content from a domain for a minimum acceptable level of Categorical Concurrence to exist between the domain and the assessment. The number of items/points, six, is based on estimating the number of items that could produce a

reasonably reliable subscale for estimating students' mastery of content on that subscale. Of course, many factors must be considered in determining what a reasonable number is, including the reliability of the subscale, the mean score and cutoff score for determining mastery. Using a procedure developed by Subkoviak (1988) and assuming that the cutoff score is the mean and that the reliability of one item is 0.10, it was estimated that six items would produce an agreement coefficient of at least 0.63. This indicates that about 63% of the group would be consistently determined to be masters or non-masters if two equivalent test administrations were employed. The agreement coefficient would increase to 0.77 if the cutoff score is increased to one standard deviation from the mean and, with a cutoff score of 1.50 standard deviations from the mean, to 0.90. Usually, states do not report student results by domains or require students to achieve a specified cutoff score on expectations related to a domain. If a state did do this, then the state would seek a higher agreement coefficient than 0.63. Six items are often assumed as a minimum for an assessment measuring content knowledge related to a domain, and as a basis for making some decisions about students' knowledge of that content under the domain. If the mean for six items is 3.00 points and one standard deviation is equal to a one-point item, then a cutoff score set at 4.00 points would produce an agreement coefficient of 0.77. Any fewer items with a mean of one-half of the items would require a cutoff that would only allow a student to miss one item. This would be a very stringent requirement, considering a reasonable standard error of measurement on the subscale. For an adaptive assessment, five items may be used instead of six because computer adaptive tests generally provide more reliable assessments with fewer items.

Depth of Knowledge Consistency

Standards and assessments can be aligned not only on the category of content covered by each but also on the basis of the complexity of cognitive engagement required by each. *Depth of Knowledge (DOK) Consistency between standards and an assessment indicates alignment if what is elicited from students on the assessment includes the same types of cognitive demands related to what students are expected to know and do as stated in the standards.* For consistency to exist between the assessment and the domains, as judged in this analysis, at least 50% of the items corresponding to a domain had to be at (or above, although not common) the DOK level of the corresponding standard. The 50% level, a conservative minimum cutoff point, is based on the assumption that a minimal passing score for any one domain of 50% or higher would require the student to successfully answer at least some items at or above the DOK level of the expectations within the corresponding domains. For example, assume an assessment included six items related to one domain and students were required to answer correctly four of those items to be judged proficient—i.e. 67% of the items. If three, 50%, of the six items were at or above the DOK level of the corresponding expectations, then for a student to achieve a proficient score would require the student to answer correctly at least one item at or above the DOK level of one expectation. If a domain had between 40% and 50% of items at or above the DOK levels of the expectations, then it was reported that the criterion was "weakly" met.

Interpreting and assigning DOK levels to both standards and assessment items is an essential requirement of content alignment analysis. The DOK descriptions help to clarify what the different levels represent for each subject area. Full descriptions for each subject area and additional information about the DOK framework are available at webbalign.org.

Range of Knowledge Correspondence

For domains (conceptual categories) and assessments to be aligned, the breadth of knowledge required on both should be comparable. *The Range of Knowledge Correspondence criterion is used to judge whether a comparable span (breadth) of knowledge expected of students by a domain is the same as, or corresponds to, the span of knowledge that students need to correctly answer the assessment items/activities.* The criterion for correspondence between span of knowledge for a domain and an assessment considers the number of standards within the domain with one related assessment item/task. Fifty percent of the standards for a domain must have at least one related assessment item for the alignment on this criterion to be judged acceptable. This level is based on the assumption that students' knowledge should be tested on content from over half of the domain of knowledge for a conceptual category. This assumes that (unless otherwise specified) each expectation within a domain should be given equal weight. For this analysis, equal weight was applied to parallel MCAS. For MCAS, neither the standards documents nor the assessment specifications provided any weighting for or emphasis on any particular standard(s) within each defined domain. Note, however, that schools participating in the MCIEA PoP identified prioritized "power standards." Depending on the balance in the distribution of items and the need to have a low number of items related to any one expectation, the requirement that assessment items need to be related to more than 50% of the expectations for a domain increases the likelihood that students will have to demonstrate knowledge on more than one expectation per domain to achieve a minimal passing score. As with the other criteria, a state may choose to make the acceptable level on this criterion more rigorous by requiring an assessment to include items related to a greater number of the expectations. However, any restriction on the number of items included on the test will place an upper limit on the number of expectations that can be assessed. Range of Knowledge Correspondence is more difficult to attain if the content expectations are partitioned among a greater number of domains and if there are a large number of expectations. If 50% or more of the standards for a domain had at least one corresponding assessment item, then the Range of Knowledge criterion was met. If between 40% and 50% of the standards for a domain had a corresponding assessment item, the criterion was "weakly" met.

Balance of Representation

In addition to comparable depth and breadth of knowledge, aligned domains and assessments require that knowledge be distributed equally or proportionally in both (unless otherwise specified in test design). The Range of Knowledge criterion only considers the number of expectations with at least one assessment item within a domain; it does not take into consideration how the assessment items/activities are distributed among these expectations. *The Balance of Representation criterion is used to indicate the degree to which one standard is given more emphasis on the assessment than another.* If test design does not include constraints for balance or otherwise specify emphases, then an index is used to judge the distribution of assessment

items. This index only considers the expectations for a domain or conceptual category that have at least one corresponding assessment item. The index is computed by considering the difference in the proportion of expectations and the proportion of items assigned to the expectation. An index value of 1.00 signifies perfect balance and is obtained if the corresponding items related to a domain are equally distributed among the assessed expectations for the given domain. Index values that approach 0.00 signify that a large proportion of the items assess only one or two of all of the expectations that were measured. Depending on the number of expectations and the number of items, a unimodal distribution (most items related to one expectation and only one item related to each of the remaining assessed expectations) has an index value of less than 0.50. A bimodal distribution has an index value of around 0.55 or 0.60. Index values of 0.70 or higher indicate that items/activities are distributed among all of the assessed expectations at least to some degree (e.g. nearly every assessed expectation has at least two items) and is used as the acceptable level on this criterion. Index values between 0.60 and 0.70 indicate the Balance of Representation criterion has only been "weakly" met.

Source of Challenge

This criterion is used to identify items for which the major cognitive demand is inadvertently placed and is other than the targeted domain or expectation (i.e., construct irrelevance). Bias and sensitivity issues, as well as technical issues and error, could all be reasons for an item to have a Source of Challenge problem. Such item characteristics may result in some students not answering an assessment item, answering an assessment item incorrectly even though they possess the understanding and skills being assessed or answering an item correctly but for the wrong reasons.

Appendix B
Subgroup Comparison for Comparability Analysis

Table B1. Comparison of Educator and Double-Blind Scores by Demographic Group

Category	N or Sample Size	Portfolio	MCAS	% Agree	Correlation
Hispanic	55	2.40	2.22	60.00	0.34**
Not Hispanic	113	2.80	2.42	51.33	0.40**
African American	6	2.78	2.67	NA	NA
Not African American	162	2.67	2.35	53.70	0.37**
Native American	11	2.41	2.55	81.82	0.71**
Not Native American	157	2.70	2.34	52.23	0.37**
Asian	15	3.09	2.40	60.00	0.53**
Not Asian	153	2.65	2.35	53.59	0.37**
Not White	69	2.51	2.25	53.62	0.33**
White	99	2.75	2.43	54.55	0.40**
Female	87	2.79	2.53	50.57	0.41**
Male	81	2.58	2.17	58.02	0.27**
FRL	79	2.40	2.24	55.70	0.39**
No FRL	89	2.90	2.46	52.81	0.38**
Not SpEd	137	2.79	2.42	54.74	0.37**
SpEd	31	2.11	2.09	51.61	0.35**
EL	26	2.04	2.04	53.85	0.30
Not EL	142	2.75	2.42	54.23	0.38**

**$p < .01$

Table B2. Comparison of Educator and MCAS Scores by Demographic Group

Category	N or Sample Size	Portfolio	MCAS	% Agree	Correlation
Hispanic	169	2.34	1.98	40.83	0.43**
Not Hispanic	378	2.74	2.53	45.50	0.50**
African American	35	2.76	2.50	51.43	0.61**
Not African American	512	2.59	2.35	43.55	0.50**
Native American	37	2.37	1.82	40.54	0.59**
Not Native American	510	2.62	2.40	44.31	0.49**
Asian	33	2.90	3.03	48.48	0.45**
Not Asian	514	2.58	2.32	43.77	0.50**
Not White	158	2.39	2.32	45.57	0.55**
White	389	2.71	2.37	43.44	0.48**
Female	266	2.70	2.42	43.61	0.53**
Male	281	2.52	2.30	44.48	0.47**
FRL	241	2.33	1.97	39.83	0.46**
No FRL	306	2.87	2.67	47.39	0.45**
Not SpEd	458	2.73	2.49	45.63	0.46**
SpEd	89	2.06	1.71	35.96	0.44**
EL	48	1.99	1.45	31.25	0.28**
Not EL	499	2.70	2.44	45.29	0.48**

**$p < .01$