

# A TRICKY BALANCE: THE CHALLENGES AND OPPORTUNITIES OF BALANCED SYSTEMS OF ASSESSMENT

Scott Marion, Jeri Thompson, Carla Evans, Joseph Martineau, & Nathan Dadey<sup>1</sup>

National Center for the Improvement of Educational Assessment

September 24, 2018

## Introduction

The seminal publication *Knowing What Students Know: The Science and Design of Educational Assessment* (National Research Council, 2001) crystalized the call for balanced systems of assessment:

Assessments at all levels—from classroom to state—will work together in a system that is comprehensive, coherent, and continuous. In such a system, assessments would provide a variety of evidence to support educational decision making. Assessment at all levels would be linked back to the same underlying model of student learning and would provide indications of student growth over time. (p. 9)

Many authors since have helped advance the conceptualization of assessment systems put forth in National Research Council (NRC, 2001) (e.g., Coladarci, 2002; Gitomer & Duschl, 2007; Gong, 2010; Perie, Marion, & Gong, 2009; National Research Council [NRC], 2004, 2006; Shepard, 2000; and Stiggins, 2006, 2008). While the practical work on systems of assessment receded to the background during the dark days of No Child Left Behind (NCLB), it slowly returned to the fore in response to concerns about the testing regimes implemented during NCLB when stakeholders sensed an opportunity as Congress prepared to reauthorize the nation’s main education law. Many scholars continue to advance our understanding of what constitutes a well-functioning system (e.g., Chattergoon & Marion, 2016; Conley, 2014; Council of Chief State School Officers, 2015; Darling-Hammond, Wilhoit, & Pittenger, 2014; Darling-Hammond, Herman, & Pellegrino, 2013; Gong, 2010; National Research Council, 2014). Still, it has been

---

<sup>1</sup> We are grateful to Ted Coladarci and Chris Domaleski for their very helpful comments and suggestions. Any errors, however, are our own.



almost 20 years since the publication of *Knowing What Students Know*, and there are few examples of well-functioning systems, particularly systems incorporating state summative tests and assessments at other levels of the system (e.g., district, classroom). Why? In spite of recent efforts to articulate principles of assessment systems, creating a balanced assessment system is really hard!

This call for balanced assessment systems resulted from a recognition that most assessments poorly served the primary purpose of assessment: improving learning and instruction. Educators understand that large-scale summative tests are far too distal from instruction, at the wrong grain size, and administered at the wrong time of year to make a difference in their daily practice. Further, many district leaders turned to commercially available district assessments that do not clearly link to other levels of the system (Perie, Marion, & Gong, 2009). Therefore, the calls to balance assessment systems—actually *rebalance* these systems—were motivated by the desire to enhance the utility of assessments for improving learning and instruction as well as for monitoring, accountability, and evaluation.

At the Center for Assessment, we have learned much about designing and implementing high quality assessment systems over the past 20 years. In this paper, we leverage the lessons of the past to forge an ambitious agenda for ways to more thoughtfully design systems of assessment that enhance equitable learning and life opportunities for all students. To do so, we first review key conceptual issues regarding assessment system design and implementation. We then examine likely reasons why there are so few balanced assessment systems in practice.

We identify many challenges or barriers—acting alone or in concert—that arguably prevent high-fidelity implementation of balanced assessment systems. We discuss each of these challenges to better understand why each has hindered assessment system implementation. By dissecting each challenge and beginning to identify high-leverage strategies for successful implementation, we hope to help others better address—and possibly avoid—these obstacles. Before considering these challenges, we begin with an overview of balanced assessment systems: definitions, criteria, and system components. We acknowledge that overcoming any one of these challenges will be difficult at best. Therefore, we conclude with an agenda for research

Center for Assessment. *Systems of Assessment* (9/24/18).



and practice that, we believe, holds promise to advance the field so that we see more balanced assessment systems used to promote student learning.

## **Would We Know a Balanced Assessment System If We Saw It?**

### Criteria for Balanced Assessment Systems

Assessment systems are balanced when the various assessments in the system are *coherently* linked through a clear specification of the learning targets, they *comprehensively* provide multiple sources of evidence to support educational decision-making, and they *continuously* document student progress over time (NRC, 2001). These properties—coherence, continuity, and comprehensiveness—create a powerful image of a high-quality system of assessments. Further, we find that *utility* and *efficiency* also are helpful considerations when working with district and state leaders (Chattergoon, 2016; Chattergoon & Marion, 2016).

### Coherence

A coherent assessment system must be compatible with how student learning is expected to progress in a domain. An assessment system is *vertically coherent* when there is compatibility among the models of student learning underlying the system’s various assessments (NRC, 2006). We generally think of vertical coherence among assessments that range from the classroom to the state level, but we should be concerned about vertical coherence even among classroom assessments serving various purposes (e.g., grading, formative feedback). *Horizontal coherence* is the alignment among curriculum, instruction, and assessment regarding a common vision of learning and how students develop proficiency in a content domain (NRC, 2006).

Further, learning progressions arguably should serve as the organizing framework for connecting the various assessments and learning activities in a vertically coherent system (e.g., Shepard, Penuel, & Pellegrino, 2018; Wilson, 2018). Shepard et al. (2018) call for “curricular specificity” here, which is an important contribution insofar as curriculum and the associated assessments “are the means by which theories of learning come to be enacted in classrooms and potentially could be made coherent across levels of the system” (p. 3).

Both vertical and horizontal coherence are necessary for assessment systems to be balanced, but as we discuss later, horizontal coherence is difficult to achieve with state-level systems. This is because most states do not have a common vision of learning across school districts through shared curriculum or learning progressions. Content standards do not have the specificity needed to fill this void.

Comprehensiveness

*Knowing What Students Know* referred to comprehensive in terms of providing a variety of evidentiary sources to inform educational decision making. While this is an important goal, balanced systems of assessment are needed to serve the needs of multiple, and often diverse, stakeholders. Therefore, the comprehensiveness criterion should be expanded to consider how well the system’s assessments serve the needs of the multiple stakeholders, generally by relying on a range of measurement approaches in support of various educational needs. In this expanded view, meeting the comprehensiveness criterion requires that we consider the different stakeholders and contexts of the systems. Table 1 presents some commonly cited purposes and uses, along with the corresponding stakeholders and contexts.

Educational measurement professionals often remind stakeholders that assessments only serve a single purpose, or narrow set of purposes, well: “Ironically, the questions that are of most use to the state officer are of the least use to the teacher” (NRC, 2001). Therefore, meeting the comprehensiveness criterion generally means employing multiple assessments to serve the needs of the various stakeholders. This is where designers need to be particularly careful to avoid producing a chaotic set of assessments that, in the end, resembles a system no more than a pile of bricks resembles a house (Coladarci, 2002).

Table 1. Typical purposes and uses of assessments.

<b>Purposes and Uses</b>	<b>Stakeholders and Contexts</b>
Supporting instruction and learning	Teachers and students within classrooms
Grading and reporting	Teachers/students within classrooms; parents



	and principals at the school level
Supporting program/curricular evaluation	Principals/teachers at school level; curriculum/assessment leaders at district level
Monitoring trends and evaluating equity	District and school leaders; state education leaders and policy makers
Providing data for accountability	State education leaders and policy makers; district leaders

### Continuity

Continuity is the degree to which the system’s assessments provide information that allows for monitoring and evaluating progress. This activity can be conducted at the student level or at higher levels of aggregation (e.g., schools). Many have heard Al Beaton’s (1990) aphorism in reference to the National Assessment of Educational Progress (NAEP): “When measuring change, do not change the measure” (p. 165). While this advice is impossible to heed within the context of classroom assessments, it nonetheless is a good reminder of the challenges associated with measuring progress. For most purposes, it is not necessary to be as precise as NAEP, but it is still important to ensure that real change, not modeling error, is communicated.

A prominent challenge for large-scale summative assessments is to produce score information that is explicitly tied to the specific content and skills students are expected to learn. Those of us who find meaning in numbers can accurately interpret whether score differences are large or small. We acknowledge the probabilistic nature of scores and the associated general performance-level descriptions. However, even measurement specialists generally cannot interpret the results of large-scale assessments in terms of where a student is located along a trajectory from fragile to deep understanding in a particular domain (this is true whether or not assessments are vertically scaled). Briggs and Peck (2015) proposed a learning-progressions approach for grounding interpretations of both achievement and growth in terms of a student’s location along a learning continuum. Closer to the classroom, some researchers are working with educators to create assessments based on learning progressions for documenting content-referenced growth (e.g., Shepard, et al., 2018; Wilson, 2018).



## Utility

Utility is the degree to which the assessment system provides the information necessary to support its multiple and often diverse purposes. Utility is not evaluated in the abstract, but, rather, follows from a well-articulated theory of action specifying the system's intended outcomes and, in turn, the processes and mechanisms by which these outcomes are realized (e.g., Hall, 2015). To be sure, assessments are validated for specific purposes and uses. But when considering utility, we must reach beyond the score inferences that are the focus of validity evaluations. With assessments purportedly designed to improve learning and teaching, these aims often include, say, providing feedback for identifying and adjusting misunderstandings, promoting deeper learning, fostering student engagement, and perhaps enhancing self-regulation or/and related skills. Thus, utility should be evaluated by examining the extent to which each assessment experience, and the system as a whole, supports the overarching aims.

Utility requires a thoughtful articulation of the intended goals of the system and, further, a theory of action regarding how these goals are realized. In other words, it is not enough simply to announce that an assessment will improve learning and teaching. Rather, stakeholders must understand—and clearly communicate—how the proposed assessment, or set of assessments, will effect desired changes in teaching and learning. For example, will assessment results have the appropriate grain size, connections to the enacted curriculum, and timeliness so educators can act on these results? Such considerations have not been addressed sufficiently in the design of assessment systems, which is why we add utility as a criterion for balanced assessment systems.

## Efficiency

We also add the criterion efficiency. By this we mean getting the most out of assessment resources and, further, eliminating redundant, unused, and untimely assessments. Efficiency determinations identify and reduce assessments that are not serving the stated purposes or are redundant with other, more useful assessments. Unfortunately, many district personnel assume a set of assessments constitutes a system if the set contains at least summative, interim, and formative components.

### Components of a Balanced Assessment System

Our discussion so far provides the initial steps for evaluating high-quality systems of assessment. Note that we did not mention the need to select certain forms or types of assessments to comprise a system. In other words, discussions of assessment systems in the more popular literature often indicate that balanced assessment systems include summative, interim, and formative assessments. Shepard (in press) argues that formative assessment should be regarded as being part of the classroom instructional system, not the assessment system (also see Sadler, 1989, and Heritage, 2010). This view makes sense: For formative assessment to be formative, it must be inseparable from instruction. Formative assessment can be thought of as a bridge between instruction and classroom assessment. The rest of the classroom assessment system—including unit-based performance tasks, extended projects, more-traditional tests, and so on—should be coherent with the formative assessment processes in that all focus on shared learning targets.

We turn now to interim assessments, relying on the definition offered by Perie, Marion, & Gong (2009):

Assessments administered during instruction to evaluate students' knowledge and skills relative to a specific set of academic goals in order to inform policymaker or educator decisions at the classroom, school, or district level. The specific interim assessment designs are driven by the purpose and intended uses, but the results of any interim assessment must be aggregable for reporting across students, occasions, or concepts (p. 6).

Many believe that interim assessments should be part of a balanced assessment system, a notion likely fueled more by commercial vendors' advertising and marketing claims than anything else. In fact, many commercial interim assessments arguably distract educators from rich assessment opportunities and, further, threaten system coherence (as we discuss later). Thus, it is far from clear that balanced assessment systems need interim assessments to be balanced (Konstantopoulos, Miller, van der Ploeg, & Li, 2016; Li, Marion, Perie, & Gong, 2010).

Most discussions of state- or district-level assessment systems usually include at least a summative component and various types of classroom assessments. But as our discussion of utility suggests, the components of a system are determined by the system's intended purposes

Center for Assessment. Systems of Assessment (9/24/18).



and uses. That said, the state summative assessment—because of its prominent role in accountability and reporting functions—typically plays a disproportionate role in most assessment systems and is responsible for much of the system imbalance we see today. When leaders discuss district or classroom assessment systems, state summative assessments are usually considered, even if that means figuring out how to work around the influence of such assessments. Additionally, “summative” does not refer to state-level tests solely, most district and classroom assessment systems include a summative component (e.g., for awarding grades or making competency determinations).

Even though this section is titled, “Components of Balanced Assessment Systems,” most readers will recognize that we did not name specific assessment system components. It is not just that we are waffling; rather it is that system components cannot be named in the abstract. System designers need to rely on a well-specified theory of action to ensure that the various components meet the needs of the various users and uses. Such a theory of action should be created in a way to allow designers to examine the assessment system criteria discussed above.

### Systems within Systems

We know in the natural world that cellular systems reside within organs and organisms. Systems of organisms make up populations and, along with considerations of abiotic and other factors, constitute ecosystems. We are familiar with the concept of systems nested within subsystems, which are defined by their boundaries and the capacity to maintain homeostasis or equilibrium. As conceptualized in *Systems for State Science Assessment* (NRC, 2006):

- systems are organized around a specific goal;
- systems are composed of subsystems, or parts, that each serve their own purposes but also interact with other parts in ways that help the larger system to function as intended;
- the subsystems that comprise the whole must work well both independently and together for the system to function as intended;
- the parts working together can perform functions that individual components cannot perform on their own; and
- a missing or poorly operating part may cause a system to function poorly, or not at all.

Unfortunately, much of the discussion of assessment systems assumes that a state-led assessment system with district, school, and classroom components is the only model. We later discuss the challenges of developing and implementing a balanced assessment system at the state level. Shepard et al. (2018) and Marion (2018) argue that districts should be the controlling agent in the design of balanced assessment systems, and Heritage (2010) and Shepard (in press) focus on the coherence of classroom assessment systems. We address multiple layers of systems in this paper, recognizing the potential power of district and classroom balanced assessment systems. But we also suggest ways to improve the coherence and balance of state-level assessments.

### **Barriers to Assessment System Design and Implementation**

As noted above, there are few examples of balanced assessment systems in practice, even though *Knowing What Students Know* is almost 20 years old. In his nationwide search for exemplary systems of assessment, Conley (2018) found only partial systems at best. We have examined much of the relevant literature over the past 20 years, and we see little attention to the reasons why, in practice, there are so few balanced assessment systems. There are more potential barriers than we reasonably can consider here, but, in view of the research literature and our extensive experience, we believe the critical factors are the:

- the influence of politics, policy, and political boundaries on decisions pertaining to assessments;
- the influence of commercialization and proliferation of assessments;
- the lack of attention to curriculum and learning in the design of assessment systems; and
- the lack of assessment literacy at multiple levels of the system.

#### Politics and Policy

The challenges of assessment system design across political and ownership boundaries remain largely unaddressed. This is not surprising insofar as measurement and assessment researchers are not necessarily trained in policy or steeped in politics. Rather, such researchers consider how various components of the system should be designed to fulfill the needs of various stakeholders

or posit how information flows through a system (Chattergoon, 2016). However, different (and disconnected) political entities control different levels of the educational system and the corresponding assessments. This is particularly true in the U.S., but it likely is true in other decentralized contexts as well. In this section of the paper, we explore how an understanding of these political and policy issues can inform our strategic efforts to implement coherent, useful, and efficient systems of assessment.

A major issue with developing a balanced assessment system is determining who is in control. Most states are local control (some more than others). Consequently, the state controls the state end-of-year assessment, but little else. In many states, any additional state-implemented assessment is seen as an assault on the local control of curriculum (e.g., the Partnership for Assessment of Readiness for College and Careers attempt at “through-course” assessment). There is considerable rhetoric regarding the need for state-led comprehensive systems, and the state has the understandable need to control its accountability test. However, districts want to control certain district-wide assessments, and schools lay claim to even finer-grained assessments. Importantly, teachers are responsible for most classroom assessments, in service of the instructional needs of their students. Implementing balanced assessment systems cannot be a state-driven enterprise, and these political and ownership boundaries cannot be ignored.

Districts are the appropriate organizational level for developing balanced systems of assessment (e.g., Shepard et al., 2018, Marion, 2018). States generally are the wrong entity for doing this work, but they do have an important role in supporting high-quality assessment systems. Depending on the district/school relationships, district offices tend to have at least a say in many assessment decisions. There is no question that an onerous state assessment (and accountability) system can negatively influence a district’s capacity to implement a high-quality assessment system, yet the latter could serve as a buffer to a weak state system. Unfortunately, most district assessments are a poorly articulated mix of legacy assessments and “multiple measures” cobbled together into an overwhelming and often incoherent picture of student learning.

### States have a role: Tight and loose coupling

We now highlight ways that states can play a meaningful role in supporting systems of assessment. The criteria for balanced assessment systems, discussed above, reflect a tightly coupled system: information flows among the various components, from the statehouse to the classroom, to maximize efficiency and utility. This is a high bar, indeed, and likely is beyond the reach of most educational systems. In contrast, recent work on designing assessments to evaluate student learning of the Next Generation Science Standards (NRC, 2014; Marion & Penuel, 2017) brings loosely coupled systems into the discussion. Such systems have multiple levels of assessments—generally the state summative assessments and modular interim assessments—all tied to the same learning targets and vision of learning science. However, because the information would not be shared across levels of the system, such loosely coupled systems are not as efficient as ones where information from one level (e.g., classroom) could be used to support purposes at another level (e.g., accountability). That said, the benefit of loosely coupled systems is that assessment leaders must explicitly acknowledge that state tests, and perhaps interim components, should be separate from the classroom assessment systems. This may serve to stave off any unintended negative consequences of state accountability on teaching and learning, such as narrowing of the curriculum, although this is contingent on the onerousness of the accountability demands. Further, loose coupling across levels of the system clarifies that it is not for the state to fully define the components of a balanced system of assessment. Rather, it is up to district and school leaders to design and implement systems of assessments to best meet local needs.

### Turnover, or Shifting Priorities, Among Policymakers

The transient nature of educational leadership is familiar to all. Most state education chiefs have been in office for fewer than three years, similar to the average tenure of large-district superintendents. Unfortunately, the policymaker turnover rate can bring considerable shifts in policy priorities. Further, changes in political climate can make untenable what were previously acceptable policies and practices. Dealing with political differences is a formidable challenge, and we are concerned that much of educational reform is personality-driven rather being sustained through explicit frameworks. Therefore, we advocate for trying to create long-term structures such as policy documents (perhaps even legislation), long-serving and apolitical

Center for Assessment. Systems of Assessment (9/24/18).



assessment advisory committees, and significant increases in assessment expertise around the state.

### Accountability

We would be remiss if we did not discuss the often perverse effects that state accountability requirements have had on the design and implementation of balanced assessment systems (e.g., Elmore, 2004; Hargreaves & Braun, 2013). Elmore offers a convincing view regarding the effects of consequences, or stakes, on educational systems:

It is absolutely essential to understand that when policies lay down stakes on incoherent organizations, the stakes themselves do not cause the organizations to become more coherent and effective. The stakes are mediated and refracted by the organizations on which they fall. Stakes, if they work at all, do so by mobilizing resources, capacities, knowledge, and competencies that, by definition are not present in the organization and individuals whom they are intended to affect. If the schools had the assets in advance of the stakes, they presumably would not need the stakes to mobilize them. In this context, stakes make no sense as policy instruments unless they are joined in some systematic way with assistance that is designed to create the organizational assets that are required to respond to the stakes. In the absence of this kind of assistance, most schools and systems will respond within the constraints of their existing assets, which are, by definition, inadequate to respond to the task. (p. 288)

In the world of assessment system design and implementation, these accountability pressures can distract leaders from long-term strategies, such as building the formative assessment skills of teachers, and instead cause these leaders to grasp at short-term approaches such as test preparation and products that promise a quick fix. Therefore, state leaders' first responsibility in promoting balanced assessment systems should be to take a hard look at the state's accountability policies and critically examine the potential unintended negative consequences. The Innovative Assessment Demonstration Authority (IADA) under the Every Student Succeeds Act (ESSA) allows the state to reduce considerably the use of large-scale state assessments for evaluating schools and, instead, provides for innovative work without having the state Center for Assessment. Systems of Assessment (9/24/18).



assessment results control the narrative. State leaders interested in fostering balanced assessment systems need to consider some way, either through the IADA or other means, of creating space for balanced assessment systems, especially systems with a strong focus on improving learning and instruction.

### The Commercialization and Proliferation of Assessments

We discussed the ways in which working across political boundaries introduces incoherence into assessment systems. Commercialization and proliferation of assessments is another source of incoherence. This often is an outgrowth of political incoherence because actors operating at different levels of the system may feel compelled to purchase additional assessments to fill a real or perceived need without a full consideration of how such assessments fit into the overall assessment system.

Some of the assessment proliferation at the district level is a result of historical programs that maintain once-useful assessments that never seem to get retired. However, a major cause of incoherent district assessment systems is the massive increase in interim assessments during the NCLB era and continuing today (NRC, 2010; Perie et al., 2009). Districts (and states) are flooded with offers from assessment vendors promising to improve student learning. Not all of these programs are low quality and ineffective, mind you, but many are (Konstantopoulos, et al., 2016; Li, et al. 2014), particularly because these interim assessments rarely align with the enacted curriculum or other programs of improvement. Because of low cognitive demand (e.g., Li et al., 2010) and weak alignment with the local curriculum, unfortunately, the results of these assessments likely distract educators from a deeper learning agenda.

Districts, of course, do not purchase these products to waste money. Rather, they do so because they think such assessments are a critical component of an assessment system and, in a climate of scarce resources, they are attracted to any putative tool for improving performance. For the want of meaningful information from state tests, district leaders struggle to even know what is going on in their districts; they want a “handle” on within-year performance across the district. Further, there is the belief that test results from an external entity are somehow official. Needing

an official score is not a defensible reason for using interim assessments, even if the credibility of teacher-generated information is questioned in some quarters.

There has been a longstanding concern about the misleading marketing efforts by interim assessment vendors, most egregiously by appropriating the literature supporting formative assessment (Shepard, 2005; Martineau, 2004). Other misleading marketing efforts involve silver-bullet promises that the product can validly serve almost any possible purpose, ranging from informing instruction to measuring academic growth to providing national comparisons. Supporting any one of these claims is difficult enough, but supporting multiple and diverse claims with a single assessment is a fool's errand. Other common marketing claims for these assessments are that

- they are aligned simultaneously with each state's content standards and the common standards;
- they permit the precise identification of a student's academic growth, both within and across grades; and
- they produce valid and actionable subscores based on few items.

This silver-bullet phenomenon, moreover, can create a perceived need where none exists—similar to walking into a store, seeing a product, and thinking, “hey, I need that” even though you had never thought of the product before. This often plays out in feeling the need for the aforementioned official score, even though an official score had not been needed previously. But what if the needs are real? Because the interim assessments likely will not fit those needs, districts become data-rich but information-poor.

Combating aggressive marketing promises is quite a challenge. We are reminded of the adage, “don't get in a war of words with someone who buys ink by the barrel.” In short, anything we offer for countering the proliferation of commercial assessments likely will be opposed with resources and outreach far greater than ours. That said, we nonetheless provide several suggestions for addressing this challenge. First, a coherent and consistent assessment vocabulary is needed for use throughout the assessment community. But until that happens, we suggest that as district leaders engage in developing coherent district assessment systems, they begin with a

clear definition of key terms and examples based on use cases (e.g., what formative assessment is and is not). Another approach is to ask those making the silver-bullet promises to provide a detailed theory of action regarding how their product will realize the stated goals (or how the product meets a presumed need that did not exist before). Vendors will find this challenging, and weaknesses in their arguments doubtless will surface. For educators to pose such questions, however, they must be assessment-literate to know how to appraise a theory of action and understand the nuances of vocabulary. Of course, having assessment-literate school and district leaders is one of the surest ways to combat the incoherent use of commercial assessments. Finally, a public vetting system of products (e.g., as Ed Reports does for curriculum packages) would result in more honest conversations between commercial vendors and users. In fact, the Louisiana Department of Education has done just that, although not at the level of critique and analysis the state would like, but at a level that nonetheless is understood by many of its educational leaders (R. Kockler, personal communication with S. Marion). Further, our fellow colleagues at the Center for Assessment, Erika Landl and Susan Lyons, are working with Ed Reports to develop a public evaluation system for interim assessments. We are hopeful that such public evaluations will help users make better decisions as well as encourage vendors to improve the quality of their products. But that is a long hill to climb.

### Curriculum and Balanced Assessment Systems

The role of curriculum in the design and implementation of balanced assessment system is one of the main challenges emerging from the issues of political control discussed above. The through line for both vertical and horizontal coherence is a common vision of learning through an enacted curriculum, describing how students are expected to progress from fragile to deeper levels of understanding and domain competence. The absence of a common vision of learning across districts serves as a significant barrier to state-led, and even district-led, balanced assessment systems. Further, the lack of high-quality curriculum even within districts is a threat to horizontally coherent assessment systems. In fact, the lack of attention to curriculum (and perhaps learning progressions) is a considerable barrier to the design and implementation of balanced assessment systems at both the state and district levels. Below, we explore some ways these curricular barriers play out in practice and, at the district level, offer some approaches for moving forward.

## Content Standards and Curriculum

Some might argue, “but we have common content standards, isn’t that the same thing?”

Curriculum and content standards are not the same. Content standards are broad statements defining the specific learning and the general cognitive demands that students should attain by the end of a grade level or grade span. They typically outline the end goals of learning. In contrast, curriculum describes the scope or breadth of the content and the sequence for learning. Curriculum provides the specificity and organizational framework that creates coherence among the standards, instruction, and assessment. Curriculum generally consists of the knowledge and skills in subjects that teachers teach and students are expected to learn (Shepard et. al, 2018). Curriculum also includes instructional materials and resources. Teachers typically plan their instruction based on the curriculum and embedded learning targets, and they then administer assessments to measure the corresponding knowledge and skills attained.

The need for creating balanced assessment systems with curriculum as a focal point is not new (Bass & Glaser, 2004; Pellegrino, 2006; Popham, 2016; Shepard et al., 2018). Classroom and formative assessment researchers (e.g., Shepard, 2000) were among the first to emphasize curriculum as a central part of balanced assessment systems. In fact, Pellegrino (2006) noted that “unless our approach to assessment is changed substantially so that it can support processes of teaching and learning focused on deep learning and understanding” the attainment of high levels of achievement, including “adaptive expertise” or the transfer of knowledge, will not occur. Assessment systems cannot support these teaching and learning processes unless each assessment is linked closely to how students are expected to learn the content and skills.

Assessments in a coherent system cannot be curriculum agnostic; rather, they should be clearly and purposefully aligned with the learning targets and how learning is expected to progress in the specific content domain. High-quality curriculum provides the framework for designing rich and varied assessments and is the lens through which one appraises the results. Again, some might question why it is not enough to connect the various assessments in the system to the content standards. If assessments are to help reveal where students are along some progression of learning, then it is critical the assessments be designed with a clear understanding of how

Center for Assessment. Systems of Assessment (9/24/18).



students are expected to move through the domain rather than skipping from one end-of-year set of content standards to the next year.

Horizontal coherence is not all or nothing. Rather, it falls on a continuum from a tight linkage to coherence only with the end-of-year content standards. Tight coherence must be in place to support improvements in instruction and learning, so any assessments purporting to serve such purposes must meet this coherence criterion. Assessments having a program evaluation role may still serve that use if they are not as connected with the curriculum as instructional assessments, but users should clearly understand the tradeoffs in using an assessment that does not align closely with the specific curriculum. For example, if the assessment's purpose was to provide evaluation information regarding the efficacy of various curriculum packages being used in a single district, then a fair evaluation would not use a single assessment tied to a particular curriculum. Additionally, assessments serving a long-term monitoring function may be exempt from the curricular coherence requirement because, by design, such assessments purportedly transcend changes in local curriculum (e.g., NAEP).

Unfortunately, most school districts rely on purchased curriculum and programs to determine what should be taught, and how. Painstaking work conducted over the past several years by EdReports<sup>2</sup> and the Louisiana Department of Education<sup>3</sup> indicates that many commercially available curricular materials fall short in quality. For example, outdated learning theories can support a coherent instruction-assessment-curriculum system, but such a system will not support the type of learning necessary to have students develop deep understandings (Shepard, 2000). In other words, weak curriculum will perpetuate a misalignment of the cognitive and attitudinal learning valued by the district. More recently, Shepard et al. (2018) and Wilson (2018) called for engaging teachers directly in the development and use of learning progressions to serve as a foundation for curricular units and assessments.

---

<sup>2</sup> See: <https://www.edreports.org/>

<sup>3</sup> See: <https://www.louisianabelieves.com/academics/ONLINE-INSTRUCTIONAL-MATERIALS-REVIEWS>



### Overcoming (or working around) the curriculum barrier

Overcoming curriculum barriers can present opportunities for stakeholders to create a shared vision for a balanced assessment system and an understanding of the role of curriculum when designing assessments. Creating a balanced assessment system that focuses on improving teaching and learning involves more than just changing the assessments and will demand varying levels of support (Bass & Glaser, 2004; Shepard et. al, 2018). We discuss three interrelated strategies for helping to better connect curriculum, learning, and assessment:

- developing a clear vision of teaching and learning,
- engaging in curriculum and assessment mapping, and
- designing and implementing curriculum replacement units.

*Clear Vision of Teaching and Learning.* Overcoming the curriculum-agnostic barrier in developing a balanced assessment system requires districts to begin with a clear vision or theory of action of what learning is valued, including the prioritization of content and the degree to which students should be able to demonstrate their cognitive and non-cognitive abilities. District leaders will need to consider the knowledge they want students to have, but also determine how students should be expected to demonstrate that knowledge. This vision must be grounded in an understanding of how students learn, and it must represent important thinking and problem-solving skills required in each of the content disciplines. This includes understanding that learning is active, requires self-monitoring and self-awareness, and moves beyond a mere accumulation of information (NRC, 2001; Shepard, 2000). Additionally, this vision necessitates a developmental approach to assessment: considering how students' understanding of content develops over time, with instruction adjusted to meet student needs. By developing this shared vision of teaching and learning, districts can begin to implement more challenging classroom tasks that address learning processes as well as learning outcomes—and ultimately assess students similarly, such as through performance assessments. Although these assessments may not be part of an external accountability system, they will enhance curriculum, instruction, and improve student learning (Shepard, 2000).



*Curriculum and Assessment Mapping.* Once a vision has been clarified and shared with the various stakeholders, the district should map their existing curriculum and assessments to these learning priorities, determining any gaps, overlaps, and mismatches. District educators will need to make decisions to embed missing curriculum units and assessments and, further, to eliminate unnecessary units and assessments, which may be a difficult decision to make. Many districts have legacy assessments tied to outdated purposes. For example, the district may still be administering a norm-referenced test that was first adopted for reasons no longer relevant. Additionally, educators must recognize where there is a misalignment of curriculum and assessments. For example, there will be a serious mismatch if the curriculum focuses on problem solving and reasoning, whereas the assessments measure decontextualized bits of knowledge. Through this mapping process, educators identify the summative assessments administered in the course or grade, determining factors such as

- the content focus of each assessment as a whole, considering the alignment to key standards or competencies;
- the type of assessment items on the various assessments (e.g., selected response, open-ended, performance-based), focusing on the balance of discrete content skills and performance; and
- the cognitive rigor of the assessment items and the assessment as a whole, including opportunities for an integration of knowledge and skills.

An analysis of these maps is required in order to identify the gaps and overlaps in the current assessment system, both within and across grades and content areas.

*Development of Curricular Replacement Units.* States generally do not have a say in curriculum decisions, and districts relying on commercial products cannot upgrade their existing curriculum at the snap of a finger. Most school districts are on a curriculum replacement schedule of roughly 7-10 years and less frequently, unfortunately, in the neediest districts. Instead of accepting this situation as is, districts should take the opportunity to re-vision the role that teachers and other educators can play in the curriculum, instruction, and assessment process.



There are multiple pathways for doing so. The development of curricular replacement units is one such pathway where both Thompson and Marion have had success with multiple districts. As Marion and Shepard (2010) clarify, these units are designed

to address the similar or same topics as existing units, but would do so in ways that embody the standards or expectations not currently addressed, and promote deeper learning than what typically occurs. These units replace existing units and would not be an add-on to a curriculum. (p. 1)

Well-designed curricular replacement units can eliminate surface-level practices and, further, provide the foundation for structuring instructional activities that are tied to a big idea of the discipline. Such units also inform the development of a unit-based assessment system where educators design pre-assessments, anticipate potential formative probes and observations, and create rich performance tasks for both instructional purposes and unit summative evaluations. As students engage in these unit-based tasks, whether for instructional or assessment purposes, teachers more clearly can differentiate and communicate various qualities of thinking, reasoning, and problem-solving. The teacher's learning-progression schema is more fully developed as a consequence, which contributes to better instructional decision-making and analytic task-specific assessment practices (Bass & Glaser, 2004).

Replacement units also provide a foundation for the design of a coherent set of assessments. Importantly, these units support actionable interpretation of both the formative activities and the performance tasks. By analyzing and interpreting student work through a clear and systematic process, teachers can improve their instructional decisions and student learning. (Thompson, 2014).

Developing a replacement unit is a good start, but more meaningful advances in curriculum and assessment are realized when multiple units are developed to occur throughout the school year. And this is particularly true if these units are connected to an underlying learning progression. The research-practice partnerships for developing learning progressions in support of learning and assessment are compelling testimony regarding what is possible (see Wilson, 2018, and Shepard et al., 2018). We are optimistic that curriculum replacement units, tied to expected

progressions of learning, are sufficiently tangible to provide educators with a useful strategy for moving closer to this vision.

### Assessment Literacy for Balanced Assessment Systems

Inadequate assessment literacy among stakeholders is a major barrier to the successful implementation of balanced assessment systems. Discussions of assessment literacy often center on the knowledge and skills educators need for properly designing, selecting, interpreting, and using assessments in the classroom—an important need, to be sure. When teachers do not know how to differentiate assessment quality, for example, they may use assessments found in the back of textbooks or on the Internet, without any consideration of the extent to which the assessment is gathering appropriate evidence about student learning of intended learning targets. However, the call for improved assessment literacy is not restricted to teachers.

Much of the blame for assessment system incoherence arguably falls on state, district, and school leaders, who often are the decision-makers regarding assessment choices. The implementation of balanced assessment systems requires that educators and leaders understand the features of high-quality balanced assessment systems, and at all levels: classroom, district, and state. Diverse stakeholders request information from the balanced assessment system, and they typically are motivated by different interests and purposes such as evaluating programs, monitoring trends in student learning, or improving instruction. The quality of a balanced assessment system depends on the capacity of stakeholders to use their assessment literacy to design and/or select high-quality assessments, accurately interpret the corresponding results, and subsequently make appropriate judgments and decisions. Unfortunately, administrators and policymakers often resort to ideology, preconceptions, and misleading sales pitches to make decisions (Coburn, Toure, & Yamashita, 2009; Gerzon, 2015).

Further, assessment literacy includes an understanding of how systems of assessments should be coherently linked together through a common learning model. Shepard's (1991) observation that most measurement professionals were stuck in a behaviorist paradigm is only slightly less true today. Our experience suggests that this myopia is not limited to measurement professionals. In short, if curriculum and assessment reform

Center for Assessment. Systems of Assessment (9/24/18).



initiatives are to be successful, educators and other stakeholders must be given opportunities develop contemporary understandings of how students learn.

The cry for greater assessment literacy is not new (Stiggins, 1991), with calls for corresponding improvement in both preservice and in-service teacher education (Brookhart, 2011; Stiggins, 1999). Nonetheless, there appears to be limited growth in the assessment literacy of educators. Does this mean educators are incapable of learning in this regard? Of course not. Rather it likely means we have been going about this in unproductive and possibly misguided ways.

### Assessment Literacy to Support Balanced Systems of Assessment

There are different, if related, demands for the various stakeholders to support the design and implementation of balanced systems of assessment, informed by their degree of assessment literacy. We discuss this with respect to educators, school and district leaders, and then state policy leaders.

#### Educators

Educators are critical actors if assessments are going to be used to support improved student learning. We do not question the advantages of having teachers understand how to interpret and use large-scale and interim assessment results, but we assert the highest priority must be improving the assessment literacy necessary for supporting highly useful classroom assessment systems. We agree with Shepard (in press) that professional development in assessment at the classroom level should be inseparable from efforts to support ambitious teaching practices and meaningful curricular reforms (also see Penuel & Shepard, 2016, Shepard et al., 2018).

Like Putnam and Borko (2000), we also believe that teacher learning in general, and assessment literacy in particular, operates from a situative perspective. This view eschews the provision of a single, exhaustive list of knowledge, skills, and abilities that any assessment-literate educator must possess. Rather, educators need to apply assessment concepts and principles in the particular situations they are likely to encounter in practice. Decision-making based on assessment results is complicated and often requires understanding of the larger context, and forces at play, in order to make better choices.

We are mindful of Phillip Schlechty’s charge when thinking about the assessment and learning literacy of educators: “The business of schools is to invent tasks, activities, and assignments that the students find engaging and that bring them into profound interactions with content and processes they will need to master to be judged well educated” (Schlechty, 2001).

Ultimately, educators must be able to design both instructional and assessment activities that allow students, parents, and teachers to understand the scope of student understanding relative to the intended learning processes and outcomes. Ultimately, educators should view assessment as the process of reasoning from evidence (NRC, 2001), and that in doing so, they learn more about their students, the subject matter, and how learning develops. This mindset also helps educators recognize that assessment results are mere estimates, and these estimates vary considerably in their usefulness for characterizing student performance and the consistency with which such performance can be characterized. We recognize that there is a lot packed into these aspirational ideas, so we unpack them a bit below.

To design high-quality tasks for both instruction and assessment means, educators must have a working knowledge of the desiderata for supporting meaningful interactions between students and content. This includes an understanding of cognitive complexity—what makes a task more or less complex in a specific domain. It also includes knowing how to structure tasks to elicit the desired evidence, scaffold the interactions among students, content, and educators, and ensure that tasks are accessible to all students. Educators also should be deft at evaluating student work—first descriptively, to gain insights into student thinking and task quality, and then more inferentially by developing tools for scoring student work.

Further, educators should understand the principal criteria for balanced assessment systems—coherence, comprehensiveness, continuity, utility, and efficiency—and their application in practice. Are such considerations, in the eyes of educators, too abstract and distant from classroom life? We think not. Educators regularly work with multiple measures, whether for student grading or program placement. Student performance on these multiple measures often is summarized using traditional approaches, such as simple averages, that may mask more than they reveal. These multiple measures provide the context for educators to initiate important

conversations about how, through the thoughtful design of systems of assessments, we can make more accurate and useful decisions about students.

Shepard (in press) notes that teaching and assessing in “fundamentally different ways is a complex and daunting task,” and it is misguided to believe that teachers can engage in this work alone or without significant support. Further, coherent and effective classroom assessment systems must be integrated with high-leverage teaching practices and rich curriculum. We agree, and we support the collaborative sense-making necessary through professional learning communities (PLC) and other forms of cross-teacher engagement. However, we doubt these person-to-person approaches can support reforms at the scale necessary to be successful.

Rather, we find in our work that assessment literacy can be improved at scale by using tools from a sociocultural perspective, particularly Lave and Wenger’s (1991) concept of “legitimate peripheral participation,” where apprentices learn to be masters. We have helped several states and school districts use this tool to build cadres of local assessment experts, who, in turn, ensure that the enhanced assessment learning is sustained. Developing an effective cadre of experts requires ongoing professional development as well as ample opportunity for those engaging in the work to share successes and concerns.

The sociocultural approach for building expertise is aided through the use of tools and processes to support assessment quality. Assessment/task design templates, student work analysis protocols, and tools for assessment quality review all provide educators with resources they can continue to use in PLCs and other collaborative-learning contexts.

Principled assessment design approaches (e.g., Misley, Steinberg, Almond, 2003; NRC, 2001) are reshaping large-scale assessment in disciplined and positive ways. We have adapted Mislevy’s Evidence Centered Design (ECD) framework for use with teams of educators in collaboratively designing rich performance tasks (Marion & Landl, 2017).<sup>4</sup> A principled

---

<sup>4</sup> We have developed a full slate of tools and templates to help educators work through a modified ECD process. This work has occurred largely in NH’s Performance Assessment of Competency Education (PACE) project, but



assessment design approach connects student learning with assessment design. This iterative approach begins by having teachers identify the claims they want to make about student performance through the careful unpacking of the knowledge and skills students are expected to demonstrate. Educators then engage in a thought experiment about the nature of evidence that would demonstrate to them that a student indeed had mastered the intended learning targets. Finally, teachers design curriculum-embedded assessment tasks that arguably elicit the needed evidence. This is a developmentally important exercise, to be sure, for teachers must apply their assessment literacy in a decidedly authentic context.

Student work analysis protocols generally take two forms. First, such protocols provide information regarding how well the assessment task elicits the desired evidence. A second approach is to use protocols that reveal student thinking that allows educators to describe the evidence related to students' demonstrations of their learning. These insights into how students learn and progress, developed through the interrogation of assessment evidence, support teachers' understanding of assessment, learning, and instructional planning.

Educators also must be able to determine assessment quality when selecting assessments. For example, educators could be given an assessment review tool for evaluating the quality of a performance assessment with respect to alignment, cognitive complexity, fairness, accessibility, text complexity, and scoring guidelines and criteria. In our experience, educators quickly realize that their assessments typically fall short in probing students' depth of knowledge and, instead, dwell on low-level knowledge and skills. This realization creates an important cognitive dissonance between the deeper learning goals that educators' espouse and what their local assessments actually measure—important because it can result in greater self-consciousness as one designs or selects assessments.

---

also with Alabama science educators. All of these materials will be posted shortly in the soon-to-be-released Center for Assessment Performance Assessment Toolkit.



### School and district leaders

School and district leaders figure prominently in the design of balanced assessment systems. They should be leaders in the design of district assessment systems, but they also must understand the hard work required of teachers. Much of the discussion of assessment literacy focuses on the teacher; there is considerably less attention devoted to helping principals and central office personnel become assessment leaders. Addressing the latter is important because district leaders, in particular, are responsible for selecting interim and other commercial assessments, which, as we saw above, may cause considerable incoherence in district assessment systems.

Like educators, school and district leaders must have a firm understanding of the design and implementation of balanced systems and the corresponding criteria. They certainly should have the assessment literacy required to evaluate the quality of individual assessments. Perhaps most importantly, however, school and district leaders must understand how to facilitate adult learning and establish a learning culture in their schools.

School and district leaders need tools such as assessment audits to help them evaluate their existing collections of assessments to begin the work of designing well-functioning systems of assessment. For example, district leaders can use the Student Assessment Inventory for School Districts<sup>5</sup> for examining their assessment systems. Similarly, the assessment mapping technique, discussed earlier, includes all of the summative assessments given within a grade and subject area over the course of the year, mapped to the district competencies and/or state content standards. While assessment audits and assessment mapping provide only a high-level view of local assessment systems, leaders and teachers alike can use these tools' results for framing questions about the degree to which the set of assessments satisfies the comprehensiveness criterion for balanced assessment systems.

---

<sup>5</sup> Achieve's *Student Assessment Inventory for School Districts* and related resources can be found at [www.achieve.org](http://www.achieve.org).



As educators review their local assessment systems, they can ask larger questions about the utility of each assessment. For example, does each assessment provide useful information for deepening student learning; improving instructional quality; and supporting administrators in making better decisions about curricular resources, programs, or personnel? Does the K-12 assessment system promote a common vision of teaching and learning, and does it engender more student agency over time? If not, a regular review cycle provides the important opportunity for teachers and administrators to consider, in collaboration, how to improve the assessment system's coherence, utility, and efficiency.

School and district leaders also can use the processes described above to begin evaluating the quality of commercially available products. This can help leaders distinguish between product marketing with actual quality. For example, when marketing materials that promise interim/benchmark assessments will serve all possible purposes, leaders need a framework for evaluating such claims. The assessment review tool is helpful here, but such technical work requires a more in-depth review of the kind being developed by Landl and Lyons (in press) for EdReports. Conducting such a review is beyond the scope of almost all non-measurement personnel, but educational leaders need to be fluent enough with key concepts such as alignment, cognitive complexity, accessibility, and error so they can meaningfully interpret the results of such expert reviews. However, even without training in sophisticated measurement concepts, educational leaders can ask hard questions about utility. For example, they should ask whether an interim/benchmark assessment is really necessary or useful for making better educational decisions about students, programs, or personnel and if so by which processes and mechanisms will this usefulness play out?

The most important role for an educational leader is to establish a local culture of learning and assessment. We are reminded Dick Elmore's discussion of the "instructional core" (City, Elmore, Fiarman, & Teitel, 2003, p. 24):

There are only three ways to improve student learning at scale: You can raise the level of the content that students are taught. You can increase the skill and knowledge that teachers bring to the teaching of that content. And you can increase the level of students' active learning of the content. That's it. Everything

Center for Assessment. Systems of Assessment (9/24/18).



else is instrumental. That is, everything that's not in the instructional core can only affect student learning and performance by, in some way, influencing what goes on inside the core. Schools don't improve through political and managerial incantation; they improve through the complex and demanding work of teaching and learning.

Obviously, there is a lot more to creating a learning and assessment culture in schools than simply reading this paragraph to school staff. Again, our focus here is on the assessment literacy necessary for designing and productively using balanced assessment systems. As we note throughout, utility is an important criterion for assessments and assessment systems. In our experience, collaboratively examining student work, initially with expert facilitation, enables educators to more thoughtfully consider issues of utility. For example, making the examination of student work a regular part of every faculty and PLC meeting promotes important discussions of how well students are learning, how student learning is progressing over time, and how school personnel can improve that learning. Such examinations of assessment utility can help educators and leaders first describe, and then draw inferences about, the ways in which different assessments elicit desired evidence of student thinking and performance. Such discussions should lead to conversations about student learning, curriculum, equity, instruction, and other critical aspects of schooling.

### State policy leaders

Prior to NCLB, some states experimented with state-led or state-supported systems of assessment (e.g., Kentucky, Maine, Maryland, and Wyoming), attempting to bridge the gap between large-scale and local assessment systems (NRC, 2003). The high stakes associated with what many regarded to be an invalid school accountability system, along with the large increase of state summative assessments (in most states), swamped any progress made with bridging the large-scale and local assessment gap. We are encouraged by the renewed interest in state-led balanced systems of assessment, despite our skepticism that states are the appropriate locus of control for such systems. While districts, and perhaps schools, are the more appropriate loci for balanced assessment systems, states, because of federal and state accountability and assessment requirements, can have an disproportionate influence on any system operating within the respective state. State policy leaders, therefore, must have assessment literacy.

Center for Assessment. Systems of Assessment (9/24/18).



While we do not believe state policy leaders require the same type of assessment literacy as teachers, they at least should be ever mindful of the following:

- Large-scale assessment serves a rather limited uses (particularly monitoring and evaluation).
- There are no magic-bullet assessments. This surfaces in discussions of subscores (e.g., algebraic reasoning or numbers and operations within mathematics), as one example, where policy makers may push for as many subscores as possible, believing that teachers will be able to act on them (even with technical advisors arguing otherwise).
- The long-term stability of the state assessment system is critical for serving its monitoring function and to minimize confusion in districts, schools, and classrooms.
- There is a plethora of research on the negative unintended consequences associated with high-stakes accountability tests, and this evidence should be considered carefully in the formulation of any new test-based accountability policies.
- The results of any test contains uncertainty, and leaders therefore should not attribute undue importance to small differences (e.g., between groups, or from one year to the next).

Making state policy leaders assessment literate begins with establishing a clear vision of learning that goes beyond the content standards. For example, groups such as EdLeader21<sup>6</sup> have worked with states and districts in developing a portrait of a graduate,<sup>7</sup> which helps stakeholders develop a shared understanding of the knowledge, skills, and dispositions expected of all students. Once a common vision of learning is established, state policy leaders, guided by their expert staff members, can begin to outline a theory of action for how assessment and accountability supports this vision. We expect this exercise to cause productive discussions of the proper role of state-level assessment versus high-quality district- and school-level assessment, the unintended negative consequences that accountability pressures may have on assessment practices, and the

---

<sup>6</sup> <http://www.edleader21.com/home>

<sup>7</sup> <https://portraitofagraduate.org/>



importance of stability of assessment programs so that educators are not distracted from the hard work of teaching and learning.

Inadequate assessment literacy among educators, administrators, and policymakers pose significant barriers to the design and implementation of balanced assessment systems. If districts indeed are the locus of control for balanced assessment systems, then developing the assessment literacy of its educators and leaders is critical to the design and implementation of high-quality balanced systems. Similarly, given the importance of the state assessment in balanced systems of assessment, we must attend to and support increases in the assessment literacy of state policy leaders.

### **Moving to an Agenda for Research and Practice**

The challenges we have considered above perhaps make this work seem ominous. While we provided a few rays of hope in our discussion, the field has a long way to go before high-quality balanced systems of assessment are commonplace. On this 20<sup>th</sup> anniversary of the Center for Assessment, we are setting an agenda for research and practice to guide our work and, we hope, to motivate others to join us in what must be a broad-based, collaborative effort. At least four concurrent strands of work are needed to ensure progress in this regard: conceptual, practical, research and evaluation, and policy.

We consider each strand below. This agenda is a work in progress, and we invite the reader to think with us on how best to move forward.

#### Conceptual work

*Knowing What Students Know* laid out the conceptual underpinnings of balanced assessment systems. This seminal work was a great start, and others have built on it over the years (e.g., NRC, 2006, 2014), but we are still operating at about 10,000 feet. In particular, additional work is needed on certain conceptual aspects of balanced assessment systems.



## Purposes and Uses

The importance of purpose has been a prevalent theme in much of the literature on balanced and comprehensive assessment systems (e.g., Coladarci, 2002; NRC, 2001; Perie, et al., 2009; Shepard, et al., 2018). We find that purpose and use rarely are articulated in sufficient detail to guide design and interpretation. Thus, understanding specification of assessment purposes is an important area of future research. Perie et al. (2009) outlined specific uses for interim assessments that function within a comprehensive assessment system; and building on this work, we seek to demonstrate how purposes must be clearly articulated in the system design. To so do, we propose that each assessment within a system be carefully described in terms of (a) what content is covered, (b) how the content is covered (types of tasks), (c) the timing of assessment administration, and (f) how the results are to be used and by whom.

## The criteria

We also wish to examine potential conflicts between the criteria of comprehensiveness and coherence: whether purpose might push an assessment “outside” the system of assessments. For example, if a system comprises classroom-, school-, and district-level assessments that all have the announced purpose of informing instruction, then it would be questionable to consider a state-level, federally mandated accountability assessment as part of that system. Often, high-stakes accountability purposes for an assessment may compromise other purposes allocated to a specific assessment (cf., Campbell, 1979). This disconnect may be a reason, in addition to political boundaries, why it is exceedingly difficult to find systems of assessments spanning classroom to state. Moreover, disconnects in purpose also may explain why many systems of assessments are so incoherent across levels of the educational system (not only between the state and other levels, but across every other level as well).

These types of separations of potential purposes suggest that a common theory of learning (NRC, 2001) may not be enough to unify a system of assessment. Although a common theory of learning may provide continuity and coherence, the purposes for the various assessments within a system may work against each other. Although there is a common theory of learning, then, the assessments still do not function as a system. In short, system stakeholders need to examine—

Center for Assessment. Systems of Assessment (9/24/18).



collectively and deliberatively—the degree to which widely disparate purposes can be served within a single system.

Such examinations also may surface whether assessments provide contradictory information. Consider the school district that gives an assessment following a large instructional unit to determine if students are ready for the next unit in the sequence. It is possible that students deemed ready for each successive unit nonetheless do not meet the desired level of achievement on, say, a state-level, end-of-year assessment used for federal accountability purposes. This apparent contradiction could occur for legitimate reasons, even if the assessments are both aligned with the same theory of learning. One plausible interpretation is that mastering the large instructional units does not fully prepare students for mastery of the entire span of content covered by an end-of-year assessment. However, this interpretation may be drowned out by the accountability pressures arising from the purpose of the state-level test. The district-level assessments, therefore, may be viewed as having far less value insofar as the results do not agree with those of the state-level assessment—even if the district-level assessments are meant to measure learning that leads up to end-of-year mastery. Thus, the use of assessments aligned with a common theory of learning still may fall short if the purposes of each assessment are too contradictory.

We intend to partner with other assessment researchers to advance the conceptual underpinnings of balanced assessment systems by conducting a program of research that addresses the issues above. There are additional things that nag at us, but we need more time to fully articulate these challenges. For example, how much specificity is necessary to achieve vertical coherence and horizontal coherence? At what point does comprehensiveness—serving multiple purposes and uses—undermine coherence? Do we need to implement learning-progressions work at scale—and if so, how?

### Practical

The practical component of the anticipated research agenda takes several forms. Most importantly, we commit to partnering with districts and states to find opportunities for designing and redesigning systems of assessment. In keeping with the Center’s open-source ethic, another

Center for Assessment. Systems of Assessment (9/24/18).



critical aspect of this work is to develop tools and other supports for practitioners. The last aspect of our research agenda's practical component is improving the quality, depth, and breadth of assessment literacy for multiple classes of stakeholders—a tremendous undertaking, to be sure.

### Partnerships

The field needs powerful and diverse examples of high-quality assessment systems in practice as models for others wishing to engage in this work. The Center for Assessment is working with several districts of varying size, as well as many states. We commit to partnering with districts to engage in the work of design and implementation as discussed in this paper. Our goal is to develop models of balanced assessment systems tailored to each locale. We will work as collaborative design partners, but we also will vividly capture the processes, struggles, and successes so that others can learn from these experiences as well.

We see several opportunities at the state level, particularly where states are partnering directly with their school districts. This is found most commonly now when states are pursuing flexibility through the ESSA Innovative Assessment Demonstration Authority. New Hampshire's Performance Assessment of Competency Education (PACE) is one such opportunity, which includes a mix of local assessments, common performance tasks across districts, and the state summative assessment in selected grades and subjects. As the technical leads for PACE, the Center of Assessment has a bird's eye view of how this system is meeting a variety of demands. In order to pursue our agenda, we must step back and study the assessment system issues associated with PACE and, further, include this examination as part of our regular dissemination. PACE provides an important opportunity to examine how local information flows up to the state level. We anticipate engaging with at least a few additional state-district partnerships as part of ESSA and other assessment flexibility opportunities.

We also have worked with several states having loosely coupled systems, where the state procures the end-of-year summative assessments as well as interim assessments that are designed to measure the same learn targets using similar measurement approaches. We find the most promising cases are where the interim assessments adhere to a modular design: the assessment relates to specific pieces of content and skills (e.g., standards and clusters of standards) that

Center for Assessment. Systems of Assessment (9/24/18).



districts can administer as they see fit. This is contrast to a mini-summative design—the most common for interim assessments—where each test (e.g., fall, winter, and spring) is aligned with the same end-of-year test blueprint. While this approach provides some within-year growth information, it holds little instructional promise. Therefore, we intend to emphasize modular interim assessments in our work with states and districts regarding assessment design and procurement efforts. While such an approach does not meet the coherence criterion, it arguably is better than having a multitude of interim assessment options, none of which is well-aligned with the state summative exam.

### Tools and resources

Well-designed tools and resources alone will not improve the assessment literacy of the users. The Center for Assessment has developed several widely used tools,<sup>8</sup> such as the Student Learning Objective and Text-Dependent Analysis toolkits. We also have drafted a district assessment system toolkit, which needs refinement to be serviceable in a variety of districts. Further, we are working with other partners to develop an assessment evaluation and auditing tool that goes beyond what is currently available. Using such a tool is an important exercise before a district team engages with an assessment system toolkit. Finally, we are developing a performance-based assessment toolkit, drawing on our work with PACE and other entities.

We are confident that these tools, thoughtfully used, will result in higher-quality assessments and assessment systems. But we emphasize the adverb *thoughtfully*. Among other things, local context and culture must be considered in the design and implementation of a system. People, not the tool or toolkit, bring the nuanced understandings of context and culture necessary for success in this regard—highly trained users who know when, and how, to color outside the proverbial lines of the tools and templates.

### Assessment literacy

We discussed at length, above, our use of a sociocultural framework for building assessment expertise. We have been successful in these efforts, particularly when the effort is part of an

---

<sup>8</sup> See: <https://www.nciea.org/featured-resources>



initiative that matters to participants. This was the case in Wyoming’s Body of Evidence initiative, where performance tasks developed by teams of educators were used for certifying students’ readiness for high school graduation. We have observed similar efficacy in New Hampshire’s PACE program, where collaborative teams of educators develop performance tasks used both for student-competency determinations and for schools as part of their accountability systems. But this work is slow. Moreover, we do not understand yet how to do this at the scale necessary to address current and future needs. We are beginning to work with digital resources in order to build assessment literacy virtually; we are eager to see if this approach increases our reach without compromising efficacy.

We are just beginning to understand how to meet the assessment literacy needs of state policy leaders. Linn and Herman (1997) tried to address some of the standards and assessment literacy needs of state leaders with their very clear and concise *A Policymaker’s Guide to Standards-Led Assessment*, but it was only one shot. Given the rapid turnover of state chiefs and board members, we need to determine how to create long-term structural supports for improving the assessment literacy of these state leaders.

We cannot do this on our own. We should draw on our strong partnerships with the Council for Chief State School Officers (CCSSO), Education Commission of the States (ECS), the National Council of State Legislators (NCSL) and other organizations to assist us in amplifying this work. That said, we can expect that assessment literacy is part of the job description of teachers and school leaders, whereas state leaders have many other competing demands (e.g., budgeting, politics, and communication). We need a better understanding of what it means to improve the assessment literacy of state policy leaders—what they need to know and understand—and how best to accomplish this. Further, we should identify approaches for state assessment leaders to better communicate the most critical assessment issues to their chief state school officers. For example, the latter could be directed to a targeted section or passage in *A Policymaker’s Guide to Standards-Led Assessment* (Linn & Herman, 1997) or an updated version. And echoing an earlier point, digital approaches can be more productively used here as well (YouTube, podcasts, and other easy-to-use outlets).

## Research and Evaluation

We have great hopes (although tempered by years of experience) for the initiatives we propose above. We know that, absent a corresponding research and evaluation structure, many of the efforts may well be one-offs. Therefore, research-practice partnerships are necessary for documenting proposed interventions so that others may learn from the work. For example, we asserted above that loosely coupled systems will improve the coherence and utility of the interim and summative components of the system. Such assertions must be supported by evidence, with plausible rival hypotheses and potential unintended negative consequences given due consideration. This is just one example, however: Similar efforts should accompany any of the major initiatives described above.

## Policy

We have outlined the implementation challenges associated with balanced assessment systems and, in turn, the beginnings of a research and practice agenda for advancing the field. Without attending to the policy context in the design and implementation of assessments, observing high-quality assessment systems in practice will continue to be like searching for unicorns. This is particularly true for systems that feature a state component. But because of how state accountability policies influence assessment-related work in districts and schools, these policy requirements can constrain their implementation of balanced assessment systems. We now turn to accountability policy and large-scale assessment policies, focusing on quality, footprint, and stability.

### Accountability policy

All states are required to implement a school accountability system that meets, at a minimum, federal ESSA requirements. Many states choose to go beyond the ESSA requirements by adding components or rules to the ESSA-based system or running a secondary (non-federal) accountability system. While ESSA is an improvement over NCLB, there still are requirements that influence the behavior of district and school leaders. After all, this is one of the intended effects of accountability policy. But where accountability incentives distract local educators and leaders from focusing on a deeper learning agenda, we are seeing an unintended negative consequence of accountability policy. All current state accountability systems rely on data from Center for Assessment. Systems of Assessment (9/24/18).



the statewide assessment system in English language arts and mathematics for generating at least two sets of indicators: achievement and student longitudinal growth. In many systems, statewide achievement test scores are used for even more indicators than these two. Even a high-quality state assessment will exert a disproportionate weight because of its prominent role in state accountability determinations. A research and practice agenda for balanced assessment systems therefore needs to examine how accountability requirements affect the development of balanced assessment systems. Further, we propose working with policy experts to craft model policies that both meet federal requirements and allow for the development of high-quality assessment systems. The rules associated with the innovative assessment pilot program offer a potential starting point for such work.

### Large-scale assessment policies

The ways in which state assessments are designed and used can have a significant role on the potential for the development of balanced assessment systems in practice. For example, there is an extensive body of research on the negative effects that low-quality assessments have on curriculum, instruction, and student thinking, most egregiously for educationally underserved and disadvantaged students (e.g., Madaus, Russel, & Higgins, 2009).

The reaction by well-meaning measurement professionals, content experts, and policymakers has been to create rigorous, high-quality large-scale assessments. There was an explosion of this work in the decade prior to the passage of NCLB in 2001 and, more recently, with the development of the multi-state Partnership for the Assessment of Readiness for College and Careers and the Smarter Balanced Assessment Consortium. This all sounds good, indeed. And it was, in part. For example, the field learned about constructing high-quality large-scale assessments. But the field also learned about making really long tests that still could not deliver instructionally useful information to school personnel and students. This is not surprising, and it is one reason why we focus so intently on systems of assessment. But we are faced with an apparent conundrum: While we certainly do not want low-quality tests, neither do we want high-quality tests requiring a 10-hour administration for each student.

We propose studying how to reduce the footprint (i.e., the influence of the state assessment on the rest of the system) of end-of-year summative tests without reducing assessment quality, in support of balanced-system implementation. There are many avenues of such work. First, sampling students would move us away from the NCLB mentality of “every student, every item, every standard, every year.” Matrix sampling<sup>9</sup> is one such approach, where rich school-level information is produced while limiting the amount of information available to individual students beyond a total score. Matrix sampling is not all or none, and it can work with many hybrid versions that combine both matrix sampled and common portions of the test. Sampling can also be over grades, but policymakers may object if a score is desired for every student every year. Another way to reduce the end-of-year footprint is to move away from student-level subscores (e.g., numbers and operations within mathematics). Such subscores pose technical challenges, so they rarely are as useful as stakeholders and policymakers hope. If states are willing to produce only a total score for each student (i.e., no subscores), end-of-year tests can be much shorter without too much of a loss of quality. Further, districts can pair such a design with optional modular interim assessments if more, and arguably better, information is desired about particular subdomains. These are just examples: We propose studying how to optimally configure large-scale tests to provide the required information while minimizing their negative impact on balanced assessment systems.

Stability is central to any policy instrument such as a large-scale assessment or accountability program, and we have observed in our 20 years at the Center for Assessment the negative consequences of instability in large-scale assessment policies. We know many states that have had three or more state testing programs over only five or six years. Not surprisingly, local educators often respond to such instability by simply hunkering down and doing nothing (“This, too, shall pass.”). There are many reasons for these frequent changes, but most are political. In addition to enhancing assessment literacy (which entails an understanding of the need for stability in this regard), we propose working with policy experts to develop guidance for

---

<sup>9</sup> Matrix sampling, like what is used for the National Assessment of Educational Progress (NAEP), involves distributing the test items among multiple forms of the test so that each student completes only a portion of the overall test, while the school (or other unit of analysis) receives information on all of the test items administered. Computer adaptive tests, especially multi-stage adaptive tests, are a logical extension of a matrix sampling.

policymakers that ensures the stability of large-scale assessment systems. We are not opposed to regular tweaks and improvements in state assessment system; rather, completely replacing one test with another should occur infrequently (e.g., when content standards are revised).

We understand the strong influence of politics on policy development and, in turn, how policies influence the design and implementation of balanced assessment systems. We also recognize our suggestion to limit the footprint of large-scale assessments is subject to test-based accountability policies. Such policies essentially act as a multiplier: exaggerating the negative influences of ill-conceived assessment policies such as instability.

### **Conclusion**

We return to where we started. We sense a desperate need to improve the quality and usefulness of assessments. Balanced assessment systems have been proposed for meeting many needs, but we do not see enough examples of such systems in practice to serve as models for others to emulate. We named several key challenges that explain why such assessment systems are rare, and we suggested approaches for ameliorating some of these challenges. We concluded by proposing a research and practice agenda for the Center for Assessment, our colleagues, and partners in order to focus our attention on this crucial work so that we can look back after the next 20 years and see more progress than we have seen in the almost 20 years since the publication of *Knowing What Students Know*.



## References

Achtenhagen, F. (2012). *The curriculum-instruction-assessment triad*. Retrieved from [http://www.pedocs.de/volltexte/2013/8256/pdf/ERVET\\_2012\\_1\\_Achtenhagen\\_The\\_curriculum\\_instruction.pdf](http://www.pedocs.de/volltexte/2013/8256/pdf/ERVET_2012_1_Achtenhagen_The_curriculum_instruction.pdf).

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Bass, K.M. & Glaser, R. (2004). *Developing assessments to inform teaching and learning*. Los Angeles, CA: CRESST. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.550.7008&rep=rep1&type=pdf>.

Bertenthal, M. W., Wilson, M.R., Beatty, A., & Keller, T.E. (2008). *Systems for state science assessment: Findings of the National Research Council's Committee on test design for K-12 science achievement*. Arlington, VA: NSTA. Retrieved from: <https://bearcenter.berkeley.edu/sites/default/files/Wilson19.pdf>.

Beaton, A. E. & Zwick, R. (1990). The effect of changes in the National Assessment: Disentangling the NAEP 1985-86 reading anomaly. Washington, DC.: National Center for Education Statistics, REPORT NO ETS-17-TR-21. Retrieved on August 1, 2018 from: <https://files.eric.ed.gov/fulltext/ED322216.pdf>

Black, P. & William, D. (1998). *Inside the Black Box: Raising standards through classroom assessment*. Retrieved from: <http://electronicportfolios.org/afl/InsideBlackBox.pdf>.

Briggs, D. C. and Peck, F. A. (2015). Using learning progressions to design vertical scales that support coherent inferences about student growth. *Measurement: Interdisciplinary Research and Perspectives*, 13(2), 75–99.

Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice*, 30(1), 3–12.

CCSSO (2015). *Comprehensive Statewide Assessment Systems: A framework for the role of the state education agency for improving quality and reducing burden*. Washington, DC.

Chattergoon, R. & Marion, S.F. (2016). Not as easy as it sounds: Designing a balanced assessment system. *The State Education Standard*, 16, 1, 6-9

City, E. A., Elmore, R. F., Fiarman, S. E., & Teitel, L. (2003). *Instructional rounds in education: A network approach to improving teaching and learning*. Cambridge, MA: Harvard Educational Press. [see particularly, chapter 1: The Instructional Core].

Coburn, C. E., Toure, J., & Yamashita, M. (2009). Evidence, interpretation, and persuasion: Instructional decision making at the district central office. *Teachers College Record*, 111, 4, pp. 1115–1161

Coladarci, T. (2002). Is it a house...or a pile of bricks? Important features of a local assessment system. *The Phi Delta Kappan*, 83(10), 772–774. Retrieved from <http://www.jstor.org/stable/20440251>

Conley, D. T. (2018). *The promise and practice of next generation assessment*. Cambridge, MA: Harvard University Press.

Durlak, J.A. & DuPre, E.P. (2008). Implementation Matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, 41, 237-350.

Elmore, R. F. (2004). Moving forward: Refining accountability systems. In Fuhrman, S. H. & Elmore, R. F. *Redesigning accountability systems for education* (pp.276-296). New York, NY: Teachers College Press

Gerzon, N. (2015). Structuring professional learning to develop a culture of data use: Aligning knowledge from the field and research findings. *Teachers College Record*, 117, 4, 1-28.

Gong, B. (2010). *Using balanced assessment systems to improve student learning and school capacity: An introduction*. Washington, DC: Council of Chief State School Officers. <http://www.ccsso.org/Documents/Balanced%20Assessment%20Systems%20GONG.pdf>

Hargreaves, A. & Braun, H. (2013). *Data-Driven Improvement and Accountability*. Boulder, CO: National Education Policy Center. Retrieved [date] from <http://nepc.colorado.edu/publication/data-driven-improvement-accountability/>.

Heritage, M., Kim, J., Vendlinski, T., & Herman, J. (2009). From evidence to action: A seamless process in formative assessment? *Educational Measurement: Issues and Practice*, 28(3), 24-31.

Herman, J. L. (2010). *Coherence: Key to next generation assessment success*. Retrieved from <https://files.eric.ed.gov/fulltext/ED524221.pdf>.

Konstantopoulos, S., Miller, S.R., van der Ploeg, A., & Li, W. (2016) Effects of interim assessments on student achievement: Evidence from a large-scale experiment. *Journal of Research on Educational Effectiveness*, 9:sup1, 188-208, DOI: 10.1080/19345747.2015.1116031

Lave, J. & Wenger, E. (1991). Situated learning: Legitimate peripheral participation. Cambridge, MA: Harvard University Press.

Li, Y., Marion, S.F., Perie, M. & Gong, B. (2010). An approach for evaluating the technical quality of interim assessments. *Peabody Journal of Education*, 85, 2, 163-185

Linn, R. L. & Herman, J. (1997). *A Policymaker's Guide to Standards-Led Assessment*. Denver, CO and Los Angeles, CA: Education Commission of the States (ECS) and the National Center for Research on Evaluation, Standards and Student Testing (CRESST).

Madaus, G., Russell, M., & Higgins, J. (2009). *The paradoxes of high stakes testing: How they affect students, their parents, teachers, principals, schools, and society*. Charlotte, NC: Information Age Publishing.

Marion, S. (2018). *The opportunities and challenges of a systems approach to assessment*. *Educational Measurement: Issues and Practice*, 37, 1, 45-48

Marion, S. & Shepard, L. (2010). *Let's not forget about opportunity to learn: Curricular supports for innovative assessments*. Dover, NH: Center for Assessment.

Marion, S. F., & Landl, E. (2017). *Principled assessment design for the Performance Assessment of Competency Education (PACE)*. Dover, NH: National Center for the Improvement of Educational Assessment.

Marion, S. F. & Shepard, L. A. (2017, June). *Assessment literacy to support competency-based education systems and other deeper learning efforts*. Presentation as part of iNACOL's National Leadership Webinar Series.

Marion, S., King, S., Blankenship, B., & Ponce, M. (2018). *Following their lead: Some thoughts about student-led assessment*. Dover, NH: Center for Assessment.

Martineau, J. A. (2004). *The Effect of Construct Shift on the Results of Growth and Accountability Models*. (Doctoral Dissertation), Michigan State University, East Lansing, MI.

Mislevy, R. J., Steinberg, L. S. and Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1: 3–67.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press.

Center for Assessment. *Systems of Assessment* (9/24/18).



National Research Council (2003). *Assessment in support of instruction and learning: Bridging the gap between large-scale and classroom assessment. Workshop report.* Committee on Assessment in Support of Instruction and Learning. Board on Testing and Assessment, Committee on Science Education K-12, Mathematical Sciences Education Board. Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Research Council. (2006). *Systems for state science assessment.* Washington, DC: National Academies Press.

National Research Council. (2010). *State assessment systems: Exploring best practices and innovations: Summary of two workshops.* Alexandra Beatty, Rapporteur; Committee on Best Practices for State Assessment Systems. National Research Council. Board on Testing and Assessment. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Research Council. (2014). *Developing Assessments for the Next Generation Science Standards.* Committee on Developing Assessments of Science Proficiency in K-12. Board on Testing and Assessment and Board on Science Education, James W. Pellegrino, Mark R. Wilson, Judith A. Koenig, and Alexandra S. Beatty, *Editors.* Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Pellegrino, J. W. (2006). *Rethinking and redesigning curriculum, instruction and assessment: What contemporary research and theory suggests.* National Center on Education and the Economy. Retrieved from <http://www.me.umn.edu/~cliao//Pellegrino-Rethinking-and-Redesigning.pdf>.

Perie, M., Marion, S., Gong, B., & Wurtzel, J. (2007). *The role of interim assessments in a comprehensive assessment system.* Retrieved from <https://www.achieve.org/files/TheRoleofInterimAssessments.pdf>.

Perie, M., Marion, S.F., & Gong, B. (2009). Moving towards a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice*, 28, 3, 5-13.

Penuel, W. R., & Shepard, L. A. (2016). Assessment and Teaching. In D. H. Gitomer & C. A. Bell (Eds.), *Handbook of Research on Teaching* (5th ed., pp. 787–850). Washington, DC: American Educational Research Association.

Popham, W. J. (2016). *The fatal flaw of educational assessment*. Education Week. Retrieved from <https://www.edweek.org/ew/articles/2016/03/23/the-fatal-flaw-of-educational-assessment.html>.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18, 119-144.

Schlechty, P.C. (2001) *Shaking up the schoolhouse*. San Francisco: Jossey-Bass

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, 29, 7, 4-14.

Shepard, L. A. (in press). Assessment for classroom teaching and learning. Workshop on Educational Assessment as Useful and Useable Evidence. National Academy of Education and American Academy of Political and Social Science. September 13-14, 2018

Shepard, L. A. (2005). Will Commercialization Enable or Destroy Formative Assessment? Paper presented at the ETS Invitational Conference, October 10-11, 2005, New York City

Shepard, L. A., Penuel, W. R., & Pellegrino, J. (2018). Using learning and motivation theories to coherently link formative assessment, grading practices, and large-scale assessment. *Educational Measurement: Issues and Practice*, 37, 1, 21-34.

Stiggins, R. J. (1991). Assessment literacy. *Phi Delta Kappan*, 72(7), 534–539.

Stiggins, R. J. (1999). Evaluating classroom assessment training in teacher education programs. *Educational Measurement: Issues and Practice*, 18(1), 23–27. William, D., Lee, C.,

Harrison, C., & Black, P. (2004). Teachers developing assessment for learning: Impact on pupil achievement. *Assessment in Education*, 11, 49-65.

Thompson, J. (2018). *Text dependent analysis: The need for a shift in instruction and curriculum*. Dover, NH: Center for Assessment.

Wilson, M. (2018). Making measurement important for education: The crucial role of classroom assessment. *Educational Measurement: Issues and Practice*, 37, 1, 4-20.

