2

The Struggle to Implement Balanced Assessment Systems: Explanations and Opportunities

Morgan S. Polikoff, *University of Southern California*
Ethan L. Hutt, *University of North Carolina at Chapel Hill*

CONTENTS

INTRODUCTION

Balanced assessment systems have not yet been as broadly implemented as their proponents desired. Assessment systems throughout the United States are still broadly characterized by incoherence, limited instructional utility, and at best, a modest impact on student learning. The failure of balanced assessment systems to gain a foothold in the standardized testing system that has otherwise massively expanded since the 2001 publication of the National Research Council's (NRC's) *Knowing What Students Know* raises important questions: Why have the ideas behind balanced assessment systems failed to achieve substantial impact in practice? What could be done differently if the goal is achieving greater implementation and impact of balanced assessment systems?

Our answer is that the history of balanced assessment systems underscores an important lesson in school reform: technical superiority is never sufficient to ensure adoption or implementation. Despite the backing of prominent experts and organizations as well as providing a sophisticated approach to assessment, balanced assessment systems have been implemented in only a handful of places and with only limited fidelity to the vision laid out in *Knowing What Students Know*. These developments point to the importance of thinking not just about the technical merit of the *Knowing What Students Know* framework but also about the political and organizational support necessary to secure ongoing implementation. The radical transformation of a country's educational assessment system would be a difficult task under any circumstances, but is considerably more difficult in the United States due to its decentralized structures and byzantine governance that determine assessment policies. Detailing and addressing the challenges posed by the political and organizational realities of American schooling is a crucial step in identifying a possible path for the implementation of balanced assessment systems.

In this chapter, we first briefly discuss the historical context in which the idea of balanced assessment systems emerged. Next, we describe the major tenets of balanced assessment systems as originally conceived and then consider how they have evolved and been operationalized over time. Following this, we describe attempts to implement balanced assessment systems in the past 20 years, and we consider how well they have been implemented and to what effect. Finally, we offer a set of high-level explanations for why balanced assessment systems have failed to take hold, 20 years after the initial ideas were put forth.

We note at the outset that the review and appraisal of balanced assessment systems offered below is hampered by a lack of clarity and common terminology in the field. Even the authors of this volume grappled with this issue as we were collectively writing the book, devoting time to debating key terms and definitions and even whether to use the term "balanced" at all. There are also several related terms that are used at least as frequently as "balanced assessment systems," including "comprehensive assessment systems" (e.g., Brookhart, 2013) and "next-generation assessment" (Conley, 2018). These terms map onto similar—but not identical—intellectual terrain, further obscuring the meaning of any of the three terms. A lack of clarity about the extent of this overlap, as well as the terms' relationship to each other (i.e., what, if anything, is signified in using one term rather than another) has contributed to a sense that the field has failed to cohere over time.

Compounding the challenge of inconsistent terminology, which makes identifying relevant research difficult, much of the work that has been done in this area—especially work at the district level—has not been published. The lack of published research has relegated much of the most important on the ground experience with attempting implementation of balanced assessment systems to the realm of anecdotes and secondhand knowledge. Even putting aside the issue of accuracy or general applicability of these accounts, since they are unpublished, they remain largely inaccessible to interested researchers or practitioners. Thus, this review relies, at least in part, on the research that the assessment experts leading the creation this volume suggested we include, even if that research does not claim to be about balanced assessment systems per se.

CONTEXT FOR AND ORIGINS OF THE CONCEPT OF BALANCED ASSESSMENT SYSTEMS

At first glance, it seems fitting that a report like *Knowing What Students Know* would be published the same year that Congress enacted a law intended to provide the public with more, and more precise, information about "what students know" than ever before (No Child Left Behind Act, 2001). Although the ideas behind *Knowing What Students Know* and the No Child Left Behind Act of 2001 (NCLB) were developed concurrently, the two documents were created by different groups and reflect very different visions of the future of school assessment in America. The political actors supporting NCLB saw an opportunity to use the federal government's standardizing power to combat "the soft bigotry of low expectations" (Bush, 2000) through accountability driven by clear standards, annual statewide assessments, and explicit reporting and progress requirements; while the scholars behind *Knowing What Students Know* drew on research on learning and recommended a shift in "the balance of mandates … from an emphasis on external forms of assessment to an increased emphasis on classroom formative assessment" (National Research Council, 2001, p. 14). At the very moment scholars sought to make assessments that were more authentic and proximate to everyday school practices, envisioning "that assessments at all levels—from classroom to state—will work together in a system that is comprehensive, coherent, and continuous" (National Research Council, 2001, p. 9), federal policy was pulling toward more remote assessments that were more aligned to statewide standards. As the intervening two decades have made clear, the ideas from *Knowing What Students Know* have remained adrift in a political environment focused on external accountability.

Although NCLB and *Knowing What Students Know* embodied different views of assessment, both were attempts to address long-running concerns about the achievement of American students and the capacity of American schools to meet the challenges of a changing world. Indeed, *Knowing What Students Know* articulates two core concerns facing American schools at that time. First, as *Knowing What Students Know* expounds on, there was a view that what mattered in terms of educational learning had shifted profoundly during the two decades prior to its publication. Following the economic malaise of the 1970s, the American economy was transitioning from manufacturing to service jobs. Although American manufacturing represented more than one-fifth of nonfarm jobs in 1979, these jobs would never again represent such a large portion of employment (Harris, 2020). The shift from manufacturing to service was understood

to require a considerable change in what school curricula valued and assessed. The upshot, as *Knowing What Students Know* states, was that,

> To succeed in this increasingly competitive economy, all students, not just a few, must learn how to communicate, to think and reason effectively, to solve complex problems, to work with multidimensional data and sophisticated representations, to make judgements about the accuracy of masses of information, to collaborate in diverse teams, and to demonstrate self-motivation. (National Research Council, 2001, p. 22)

These changes, all associated with the rise of the information economy, were also supposed to be amplified by technological shifts involving the increased capacities of computers, the Internet, and electronic communication like email.

Second, *Knowing What Students Know* argued that American students needed to be trained differently to be successful, fitting comfortably within a longer running narrative that American schools were or had become largely ineffectual. The consequences of the perceived deficiencies of America's schools had taken on new and higher stakes during the Cold War. Schooling was no longer simply a matter of producing good citizens or providing a means of personal advancement—instead, developing the nation's human capital was now a matter of existential economic and military importance (e.g., Tröhler, 2014). This argument provided additional non-moral justification for improving educational equality: failing to develop the talents of all American youth was a waste of one of the country's most valuable resources. During the 1970s, these views about the role of schools, coupled with America's ongoing economic woes helped drive the rapid rise and proliferation of the Minimum Competency Testing (MCT) movement (Resnick, 1980). The MCT movement, which sought to improve school performance and student achievement by requiring students to pass tests to graduate from a particular grade or school, was arguably the country's first nationwide effort at test-based accountability. Thirty-five states adopted some form of MCT by 1980.

Concerns about school performance, economic competitiveness, and global competition were bolstered by the publication of *A Nation at Risk: The Imperative for Education Reform* in 1983, which added new rhetorical heft and policy aims to the national conversation (National Commission on Excellence in Education, 1983). This report asserted that "the educational foundations of our society are presently being eroded by a rising tide of mediocrity that threatens our very future as a Nation and a people," adding that "if an unfriendly foreign power had attempted to impose on America the mediocre educational performance that exists today, we might well have viewed it as an act of war" (National Commission on Excellence in Education, 1983, p. 9). Concern about the mediocrity of school performance led *A Nation at Risk* to reject a focus on minimum competency in favor of calling for educational excellence. This excellence would be achieved by increased attention to higher standards, improved curricula, and greater accountability.

The ideas espoused in *A Nation at Risk* provided the blueprint for future educational reform. The call for better, more rigorous curriculum and standards was adopted by several professional organizations. For instance, the National Council of Teachers of Mathematics published *Curriculum and Evaluation Standards for School Mathematics* in 1989, which called for a novel and more conceptual approach to mathematics instruc-

tion (National Council of Teachers of Mathematics, 1989). Similar calls to shift subject emphasis and approach occurred in history, science, and reading. Congress also created the National Council on Education Standards and Testing in 1991, which issued a report titled *Raising Standards for American Education* and endorsed "the adoption of high national standards and the development of a system of assessments to measure progress toward those standards" across the school curriculum (National Council on Education Standards and Testing, 1992, p. 8). These calls for curricular reform were paralleled by legislative efforts aimed at increasing political pressure for improved school performance. Following a summit of state governors in Charlottesville in 1989, the federal government passed a series of bills—Goals 2000: Educate America Act (1994), Improving America's Schools Act (1994), and NCLB (2001)—that were intended to incentivize states to raise expectations for student achievement, develop new and more rigorous academic standards, and establish a more regular and robust system of assessment (McGuinn, 2006).

Set among these developments, it is easy to imagine that the *Knowing What Students Know* report committee believed there was value and promise in articulating a more sophisticated, research-driven view of assessment. Indeed, the framing of the report as an effort to cast aside outdated approaches to assessment in favor of more ambitious and rigorous ones was perfectly aligned with two decades of rhetoric about educational reform. As *Knowing What Students Know* clearly identifies, a system of reform predicated on using students' demonstrated knowledge on standardized assessments to guide system-level changes is only as good as the assessments are. *Knowing What Students Know* also incorporated recent developments in research on cognition and learning (e.g., National Research Council, 1999; Resnick & Resnick, 1992), including highlighting the situated nature of understanding and the importance of cognitive schemas in shaping a person's ability to learn, recall, and apply information in new contexts—both of which underscore the need to rethink the what, how, and when of assessments (Resnick & Resnick, 1992). *Knowing What Students Know* also reflected the view that traditional assessments, such as those that used multiple choice questions and those that required students to recall basic facts without probing their cognitive processes, could never provide adequate information about student learning or competence in a curricular domain. Likewise, the prevailing view was that the inability of existing assessment systems to measure complex knowledge and skills virtually assured that test results could neither speak to the more ambitious elements of newly adopted standards nor provide sufficiently detailed accounts of student learning to guide teaching and instruction. With these prescient concerns in mind, *Knowing What Students Know* sought to chart a different course for America's system of educational assessment.

WHAT ARE BALANCED ASSESSMENT SYSTEMS?

This section focuses on the criteria for describing and evaluating balance in assessment systems as laid out in *Knowing What Students Know*. To build up to these criteria, *Knowing What Students Know* puts forth an argument about the state of assessments in U.S. education at the time of its publication, developing the conceptual justification for balanced assessment systems. Here, we summarize this argument and then present and describe the criteria and how they were operationalized.

Knowing What Students Know begins by addressing the nature of assessment and identifies assessment's three main purposes. First, there is *formative assessment*, or assessment to assist teaching and learning. Second, there is *summative assessment*, or assessment to ascertain students' level of competency. Summative assessments can be classroom-based or large-scale, although *Knowing What Students Know* focuses on large-scale assessments given the contemporaneous policy context discussed above. Third, there is assessment to evaluate programs: these are typically based on summative assessments, but instead of using the assessment to make a judgment about an individual, it is used to make a judgment about an institution or policy. The report notes that a single assessment will not be able to serve all these purposes and that there is indeed often misalignment among the various purposes (e.g., that the kinds of assessments useful for teachers' instructional decisions are typically poorly suited for evaluation).

Knowing What Students Know argues that assessment is a process of reasoning from evidence. Assessment data (i.e., students' responses to assessment prompts) provide evidence through interpretation. A chain of reasoning helps the author of the assessment determine what to measure and establish a justification for how that measurement produces evidence to address the desired goals of the assessment. *Knowing What Students Know* focuses on the "assessment triangle," which emphasizes three essential elements underlying any assessment—a model of student cognition, a set of beliefs about the kinds of observations that will provide evidence of student competencies, and an interpretation process for making sense of the evidence (see Figure 2-1). The implication of the assessment triangle is that all three components—cognition, observation, and interpretation—must support each other for the assessment to be effective. *Knowing What Students Know* further emphasized that the model of student cognition underlying the assessment triangle should extend to curriculum and instruction.

Knowing What Students Know then discusses the state of knowledge on thinking and learning and draws implications for assessment systems, echoing other contemporary accounts of assessment for policy and practice (e.g., Shepard, 2000; Stiggins, 2001). For instance, the report concludes that assessment practices are too focused on component skills and discrete knowledge and not enough on complex aspects of student achievement. The report emphasizes the mind's cognitive architecture and concludes

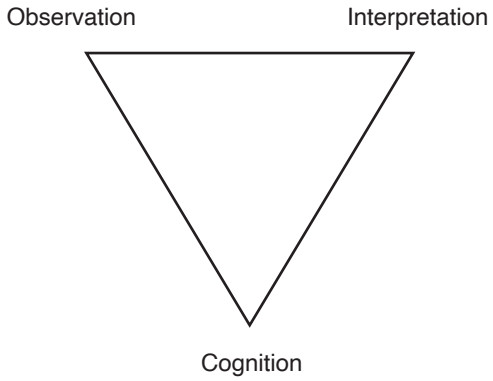

FIGURE 2-1 The assessment triangle.
SOURCE: National Research Council (2001).

that a primary focus of assessment should be on understanding the contents of long-term memory and how people use long-term memory to answer questions and solve problems. It also concludes that assessments should measure important metacognitive skills and problem-solving strategies. By making students' thinking visible, these more advanced forms of assessment can be more instructionally useful to teachers. These forms of assessment also must consider important contextual factors such as students' background knowledge, the context in which the tasks are presented, and the degree of transfer required for success in the task. Importantly though, the report also concludes that the body of evidence on creating these more advanced assessments is insufficient for practical implementation in most cases, requiring further translation to be useful in practice.

Knowing What Students Know then proceeds to describe a vision for a modern assessment system, beginning with the importance of a model of cognition and learning for assessment development. For instance, the report argues that if one is assessing to inform arithmetic instruction, test developers need to start from an understanding of how students learn in the tested domains. This model of cognition and learning should be based on empirical research in the domain, be able to differentiate between the performances of novice and expert learners, and account for variation in student learning pathways. The model should also inform assessment construction by identifying appropriate aspects of the larger theory of cognition and learning, and should lend itself to being aggregated for a variety of assessment purposes. *Knowing What Students Know* notes that there will be content areas where models of student learning are not well developed but argues that the general principles articulated above should still hold. The report goes on to discuss examples of models and their application, as well as principles of task and assessment construction for this new model and provides validation and reporting recommendations.

The Principles of Balanced Assessment Systems

With this view of assessment articulated, *Knowing What Students Know* introduces and describes the principles of balanced assessment systems. After describing features of classroom-level and large-scale assessments independently, the report advocates balancing these two forms of assessment, claiming that the status quo at the time of publication was heavily tilted toward large-scale uses. The report then introduces three principles that characterize balanced assessment systems: *comprehensiveness*, *coherence*, and *continuity*. As we discuss below, the decision to introduce principles to define balanced assessments—rather than elements or features—means that determining whether a system has achieved balance is less a categorical determination than one of degree. The challenge posed by fulfilling these principles is both technical and political. For instance, how, to what degree, and with what balance these principles should be pursued would certainly spur debate among experts. Whether the ensuing compromise would be acceptable to the public or feasible in practice given the limited time, resources, and technical expertise available in individual states, school districts, or schools is a matter likely to produce more compromise and perhaps deviation from the ideal.

With this in mind, we present the principles of balanced assessment systems and their definitions, discussing each principle in some detail with an eye toward helping readers think of them in terms of a continuum.

Comprehensiveness

The first principle of balanced assessment systems is comprehensiveness. *Knowing What Students Know* characterizes comprehensiveness by explaining:

> A *range of measurement approaches* should be used to provide a variety of evidence to support educational decision making. Educational decisions often require *more information than a single measure can provide*. As emphasized in the NRC report *High Stakes: Testing for Tracking, Promotion, and Graduation*, multiple measures take on particular importance *when important, life-altering decisions (such as high school graduation) are being made about individuals*. No single test score can be considered a definitive measure of a student's competence. *Multiple measures enhance the validity and fairness* of the inferences drawn by giving students various ways and opportunities to demonstrate their competence. The measures *could also address the quality of instruction*, providing evidence that improvements in tested achievement represent real gains in learning (NRC, 1999c)…. Further, in a comprehensive assessment system, the information derived should be *technically sound and timely* for given decisions. One must be able to *trust the accuracy* of the information and be assured that the inferences drawn from the results can be substantiated by *evidence of various types*. The technical quality of assessment is a concern primarily for external, large-scale testing; but *if classroom assessment information is to feed into the larger assessment system, the reliability, validity, and fairness of these assessments must be addressed as well*. (National Research Council, 2001, pp. 253–255, italic added by authors for emphasis)

This principle emphasizes the benefits of employing multiple assessment measures, especially in high-stakes instances. This emphasis is in direct response to prevailing uses of assessment in the years prior to *Knowing What Students Know*, especially in high stakes situations (e.g., minimum competency tests, exit exams used to award high school diplomas). *Knowing What Students Know* offers an example of a comprehensive assessment used in the United Kingdom for A-level physics, which combines multiple short sit-down assessments that include a variety of item types with laboratory exercises and essays.

According to *Knowing What Students Know*, comprehensiveness has several benefits. First, more comprehensive assessment systems provide more information than a single measure. They also enhance the validity and fairness of the inferences drawn from the data, so are therefore more trustworthy for users. Finally, more comprehensive systems may also be more instructionally valid (i.e., more useful for discerning effective and ineffective instruction). But to achieve these benefits, the comprehensive assessments must be technically sound and delivered in a timely manner, for both classroom and large-scale assessments. *Knowing What Students Know* also briefly acknowledges that comprehensiveness requires greater cost and effort in terms of assessment development, validation, and scoring.

Coherence

The second principle of balanced assessment systems is coherence. *Knowing What Students Know* defines coherence as follows:

> One dimension of coherence is that the *conceptual base or models of student learning* underlying the various external and classroom assessments within a system *should be compatible*. While a large-scale assessment might be based on a model of learning that is coarser than that underlying the assessments used in classrooms, the conceptual base for the large-scale assessment should be a broader version of one that makes sense at the finer-grained level (Mislevy, 1996). In this way, the *external assessment results will be consistent with the more detailed understanding of learning underlying classroom instruction and assessment*. As one moves up and down the levels of the system, from the classroom through the school, district, and state, assessments along this vertical dimension should align. As long as the underlying models of learning are consistent, the *assessments will complement each other rather* than present conflicting goals for learning.
>
> To keep learning at the center of the educational enterprise, assessment information must be *strongly linked to curriculum and instruction*. Thus another aspect of coherence, emphasized earlier, is that *alignment is needed among curriculum, instruction, and assessment* so that all three parts of the education system are working toward a common set of learning goals. Ideally, assessment will not simply be aligned with instruction, but *integrated seamlessly into instruction* so that teachers and students are receiving frequent but unobtrusive feedback about their progress. If assessment, curriculum, and instruction are aligned with common models of learning, it follows that they will be aligned with each other. This can be thought of as alignment along the horizontal dimension of the system.
>
> To achieve both the vertical and horizontal dimensions of coherence or alignment, *models of learning are needed that are shared by educators at different levels of the system*, from teachers to policy makers. This need might be met through a process that involves gathering together the necessary expertise, not unlike the approach used to develop state and national curriculum standards that define the content to be learned. But *current definitions of content must be significantly enhanced based on research from the cognitive sciences*. Needed are user-friendly descriptions of how students learn the content, identifying important targets for instruction and assessment (see, e.g., American Association for the Advancement of Science, 2001). (National Research Council, 2001, pp. 255–256, italic added by authors for emphasis)

Later reports explicitly distinguish and define vertical and horizontal coherence using the same concepts articulated in the excerpt above (e.g., National Research Council, 2006; Shepard et al., 2018). In the vertical dimension, a common model of student learning helps ensure that different forms and levels of assessment provide complementary, rather than conflicting, information. This definition serves as a rejoinder to large-scale assessments that were seen as poorly aligned with classroom assessments, sending teachers unclear signals about student performance and their own instructional needs. In the horizontal dimension, alignment among—or integration of—curriculum, instruction, and assessment helps ensure relevance and utility of assessment results.

The last paragraph in the definition excerpted above emphasizes the importance of learning progressions—models of student learning of content that are based in cognitive science and are widely shared and understood across the levels of the system. *Knowing What Students Know* notes that existing definitions of content, including those found in standards documents, are insufficiently linked to detailed conceptions of how students learn. In its conclusion, the report advocates for a substantially expanded research agenda to develop and test new conceptual models of student learning.

Continuity

The third and least-developed principle is continuity, which is defined in *Knowing What Students Know* as follows:

> In addition to comprehensiveness and coherence, an ideal assessment system would be designed to be *continuous*. That is, assessments should *measure student progress over time*, akin more to a videotape record than to the snapshots provided by the current system of on-demand tests. To provide such pictures of progress, *multiple sets of observations over time must be linked conceptually so that change can be observed and interpreted*. *Models of student progression in learning* should underlie the assessment system, and tests should be designed to provide information that maps back to the progression. (National Research Council, 2001, pp. 256–257, italic added by authors for emphasis)

This principle, reflecting developments in the science of learning and cognition, emphasizes models of student learning, and especially the longitudinal and temporal nature of that learning. Continuity argues for the centrality of assessments for measuring growth, as opposed to the typical one-time assessment practice. It is unstated, perhaps because this report predates the modern "value-added" movement that emphasizes the attribution of growth in student achievement to individual schools and teachers, but the implication is that the growth focus of the continuity principle refers to more than simply a statistical analysis of performance over time. Rather, the focus is on developing longitudinal models of student learning, as well as focusing assessment and reporting on student performance over time against those longitudinal models.

How Much Balance Is Enough?

One of the central tensions we return to throughout this chapter, and that we believe contributes to the difficulties in widely implementing balanced assessment systems, is the lack of clarity about how to measure balance and determine when there is sufficient balance in the system. How much balance is "enough?" As the principles above illustrate, there is no bright line or checklist that says if your system contains X, Y, and Z elements, it is sufficiently balanced.

The three principles of balanced assessment systems and the other criteria discussed later in this chapter are continuous, not dichotomous. For an assessment director seeking to create balance, the task can seem Herculean because more and different forms of assessment can always be added to make the system more comprehensive. One can always tighten the link between assessment and curriculum to bring about more coherence. And one can always add additional longitudinal measurements to deepen

continuity. In contrast to the ever-expanding possibilities of bringing "more balance" are the very real time and resource constraints for developing and implementing assessment systems. Yet, despite this challenge, we are unaware of any clear processes or guidelines for how to create sufficient comprehensiveness, coherence, and continuity. We return to this issue later, and we also note that this challenge of how much is enough is not unique to the effort to achieve balanced assessment systems. Modern argument-based conceptions of validity (Kane, 1992) also suffer from a similar challenge: how much evidence is enough to make a particular validity determination is in the eye of the beholder. Accepting that reasonable minds will differ on the technical matter of how to pursue and balance these principles in the design of an assessment system, we expect that implementation of these systems in practice would foster still more variation, further underscoring the need to assess balance along a continuum.

Updates to the Balanced Assessment Criteria

Over time, the balanced assessment criteria have been expanded and revised in various ways (Marion et al., 2019b; National Research Council, 2003, 2006, 2014). The 2003 NRC report *Assessment in Support of Instruction and Learning: Bridging the Gap Between Large-Scale and Classroom Assessment* expanded significantly on *Knowing What Students Know*, discussing the three criteria alongside two additions to the list: (1) integrated and (2) high-quality assessments (National Research Council, 2003). The 2003 report offered summaries of innovative systems at the time across the United States and the world, but also noted the lack of evidence for the effectiveness of these programs—a problem that persists today. The report also highlighted that "with a few exceptions, little effort has yet been made to transfer these programs to other settings with different characteristics" (National Research Council, 2003, p. 42).

In 2006, the NRC followed up with a deep dive into issues related to state assessment systems for science in *Systems for State Science Assessment* (National Research Council, 2006). This report also emphasized aspects of balance, focusing on horizontal, vertical, and developmental coherence in assessment systems—embodying some elements of coherence, continuity, and comprehensiveness, although using somewhat different terminology. The 2006 report helpfully described different models by which states could meet the NCLB science assessment requirements, while also raising thorny technical considerations.

By 2014, with Next Generation Science Standards (NGSS) becoming more widely adopted and the need for a more coherent approach to science assessment becoming more clear, the NRC issued a report on science assessment systems, *Developing Assessments for the Next Generation Science Standards* (National Research Council, 2014). This report brought together many of the ideas and examples discussed in previous NRC reports and included specific examples of curriculum-embedded tasks that embody the 2014 report's vision of NGSS-aligned assessment systems. The 2014 report argues that proper assessment of the NGSS *requires* elements of balance, for instance claiming that tasks must "be designed so that they can accurately locate students along a sequence of progressively more complex understandings of a core idea and successively more sophisticated applications of practices and crosscutting concepts" (i.e., continuity, National Research Council, 2014, p. 45). The 2014 report also emphasizes the need

for a systems approach to assessment of the NGSS, including classroom assessments, monitoring assessments, and indicators of opportunity to learn, while noting the difficulties associated with the lack of common curricula across jurisdictions.

Given these developments in the years since *Knowing What Students Know*, recent efforts to define balanced assessment systems introduced two additional criteria: utility and efficiency. Chapter 1 of this volume, "Reimagining Balanced Assessment Systems: An Introduction," adds another related criterion: a focus on ambitious and equitable teaching and learning, but we do not repeat that argument and explanation here.

Utility

Marion and colleagues (2019b) provide definitions of both utility and efficiency, citing their work with states to operationalize balanced assessment systems. Utility is defined as follows:

> 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 follows from a *well-articulated theory of action* specifying the system's *intended outcomes and 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 and rely on a *theory of action that spans all of the components of the system*. With assessments purportedly designed to improve learning and teaching, these aims often include: providing feedback for identifying and adjusting misunderstandings, promoting deeper learning, fostering student engagement, and/or 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*. (Marion et al., 2019b, p. 5, italic added by authors for emphasis)

This conception of utility represents another bold addition to the already ambitious balanced assessment framework. It reinforces the idea that balanced assessment systems require a coherent theory of student learning and organizational change. Importantly, this definition of utility operates at both the level of individual assessment experience and the whole system of assessment. Given the decentralized structures governing assessment in American education, we argue that utility should be evaluated at the individual district level.

While utility is to be evaluated against each individual system's theory of action, the definition sets a high bar by implying that assessment and assessment systems should not only improve teaching and learning but also "[provide] feedback for identifying and adjusting misunderstandings, [promote] deeper learning, [foster] student engagement, and/or [enhance] self-regulation or/and related skills" (Marion et al., 2019b, p. 5). If assessment systems are struggling to only improve teaching and learning, these same systems will surely struggle to achieve more complex goals. We note that while the term "utility" was not emphasized in Chapter 1 of this volume, "Reimagining Balanced Assessment Systems: An Introduction," the goal of fostering ambitious and culturally relevant instruction sets a target for the utility of district assessment systems.

Efficiency

Recognizing both the potentially boundless scope of balanced assessment systems and the growing anti-test political context of the latter half of the 2010s, a final criterion for balanced assessment was added—efficiency—which is defined as follows:

> By this we mean *getting the most out of assessment resources* and *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. (Marion et al., 2019b, p. 7, italic added by authors for emphasis)

Efficiency is a valuable criterion to thwart the "yes, and" approach to balancing an assessment system. That is, this criterion specifically forces local actors to consider whether each individual assessment in the system is necessary or superfluous, rather than simply piling new assessments on top of existing ones to achieve balance. Efficiency is also especially useful when combined with utility since efficiency focuses mainly on quantity and utility focuses on the quality of each assessment and its alignment with a theory of change.

THE IMPLEMENTATION OF BALANCED ASSESSMENT SYSTEMS OVER TIME

The assessment principles articulated in *Knowing What Students Know* reflect a new era of thinking about the role, design, and implementation of standardized assessments. Having better, more sophisticated ideas is, however, not enough to ensure their faithful adoption or implementation, even when they are backed by research and released under the auspices of an esteemed group like the NRC. Two decades later, there has been only modest success at sustained implementation of balanced assessment systems at scale. Even so, there have been several attempts to pilot new assessment systems and reorient existing systems in ways that reflect the assessment principles in *Knowing What Students Know* (in line with the observation above that balance is a matter of degree, not a bright line). In this section, we review some of these implementation efforts and consider the lessons that can be gleaned by these examples.

New Hampshire's Performance Assessment of Competency Education

Arguably the most notable example of an effort to reform an assessment system in line with the principles of *Knowing What Students Know* involves a pilot project in New Hampshire. In 2015, New Hampshire took advantage of flexibility provided by the U.S. Department of Education and received a waiver for certain elements of the testing requirements under NCLB to pilot an assessment program called New Hampshire Performance Assessment of Competency Education (PACE) (New Hampshire Department of Education, 2023). The goal of PACE was to develop an assessment system that could better serve the multiple purposes and audiences that utilize such assessment information. While annual state testing regimes developed under NCLB were providing useful information to lawmakers and the broader public, the scores produced by statewide testing were of limited value to classroom teachers in guiding

their instruction. PACE sought to address this shortcoming by developing a multi-layered assessment system involving locally developed and administered performance assessments, common assessments administered across participating districts, and the standard (Smarter Balanced) state-level assessments. Both the local and common tasks used in the PACE system are teacher-designed, and, as a result, are intended to closely resemble classroom tasks and instruction. Common tasks are developed jointly by teachers across districts and are subject to a one-year pilot testing period, during which the performance and scoring of the tasks are assessed for quality and potential biases. Following any necessary revisions, common tasks are administered across participating districts, with the same common task administered across districts in the specified grade. Local tasks are developed by individual teachers or schools, intended for local use, and are not required to undergo the same piloting or evaluation process as the common tasks. Teachers developing local tasks are encouraged to first participate in the development of a collaborative common task so that they have an opportunity to develop their knowledge about and skills for task creation. To reduce the burden on teachers to develop local tasks and to increase the number of high-quality tasks available for use, common tasks that have been piloted and made operational are added to a pool of tasks that can be drawn on by teachers and schools for use as local tasks. Depending on the grade and subject matter, local and common assessments developed under PACE are used to make the required annual state determination of student competency (Becker et al., 2017; Evans & Lyons, 2017).

The PACE pilot closely adheres to the two *Knowing What Students Know* principles of coherence and comprehensiveness. By incorporating teacher-developed competency tasks in addition to traditional standardized assessments, PACE measures of competency involve a much larger variety of tasks and, in turn, can assess a larger array of student skills and abilities than a traditional assessment system. Likewise, because the required performance assessments are so closely linked to classroom curriculum and instruction, the information provided by these assessments is much more likely to provide teachers with information that can inform their planning and instruction. This information will likely prove more useful to teachers because they are both more familiar with the local and commonly administered tasks than they would be with a standardized assessment devised by a third party but also because, at least theoretically, the preparation for and administration of the assessments provides teachers with real time feedback on their instructional practice. As the authors of a formative assessment of PACE explained, "PACE is … intended to influence instructional practices" but unlike with traditional assessments, "PACE leadership is not overly concerned about teachers 'teaching to the test.' PACE, ideally, supports 'testing to what is taught'" (Becker et al., 2017, p. xxii).

Although PACE is useful in demonstrating how the principles of *Knowing What Students Know* can be used to devise an annual statewide assessment system, the pilot project also illustrates the considerable challenges and resources necessary to implement such a system. Across the first four years of the pilot, only 14 of the state's 84 school administrative units had implemented the PACE system (Lyons et al., 2017). To participate fully in the pilot (i.e., use PACE across all available grades and subjects), districts had to commit teacher time to developing, piloting, administering, and scoring the local and common assessments required by the PACE system. Given that

full participation implicitly assumes teachers—across grade levels and subject matter expertise—have the ability to develop high quality performance tasks and assess them reliably, participation requires considerable investment in professional development for these teachers.

New Hampshire developed a three-tier system that allowed districts to build this capacity over time with appropriate state level support. Tier 3 districts, or those that had few or no classrooms implementing competency-based learning and personnel with little or no experience developing task-based performance assessments to evaluate competencies, were provided access to school-level consultants to develop local competency targets based on the state's model standards. Tier 2 districts, or those that had developed the necessary school and course level competency targets and had some—or uneven—experience developing task-based performance assessments, received access to professional development from experts on topics ranging from creating performance tasks to developing reliable scoring procedures to fostering professional learning communities. Tier 1 districts, or those that were fully implementing PACE, were provided expert consulting and coaching assistance as teachers engaged in the multi-step process of developing, piloting, revising, and implementing high quality performance tasks. Tier 1 districts also, with state assistance, invested in more advanced training for select teacher leaders focused on advanced performance assessment, including validity theory and principled assessment design; depth of knowledge, to assist with developing more cognitively demanding assessment tasks; and professional community development, to facilitate collective development of assessment tasks. Finally, teachers from Tier 1 districts were required to attend sessions during the summer where student work on PACE common tasks was discussed and scored (Lyons et al., 2017).

Producing comparable scores is an especially important component of the PACE pilot, given the goal of using local and common tasks in place of statewide assessments for some grades and subjects (Evans & Lyons, 2017). To be viable for these purposes, student scores on PACE tasks across districts needed to support the same inferences about students' knowledge and skills in a domain. Specifically, the state needed to develop scoring procedures—and processes to monitor the scoring—that would ensure that a student reported proficient on a task in one district would be rated proficient in another district. To ensure this contiguity, samples of student work must be scored by multiple teachers within and across districts to ensure inter-rater reliability and comparability of scores within and between districts. As the multiple evaluations of the PACE program cited throughout this section have found, completing these scoring processes and achieving acceptable levels of scoring reliability—the target is 60 percent—is time intensive. As one study concluded, "the practicality and feasibility of scaling up the proposed methods in a large-scale performance assessment program is a real concern particularly within a state that has many more districts or other units with a large number of different local assessment systems" (Evans & Lyons, 2017, p. 31).

A formative evaluation of PACE, likewise, found that the amount of time required of teachers to develop assessments and calibrate and score student work was an "ongoing source of remaining tension" in the pilot implementation (Becker et al., 2017, p. 49). One-quarter of the teachers surveyed disagreed or strongly disagreed with the statement that the time required by PACE was worth the benefits (Becker et al., 2017). One attempted solution was to schedule a task planning session for the weekends. While

this shift did not reduce the amount of time required, it did reduce the amount of time teachers had to be absent from their classrooms. Some districts have also experimented with reducing the amount of time required to score student work by eliminating scoring calibration sessions and shifting the work of scoring from the school year to the summer. The result, however, was that these districts had substantially lower inter-rater reliability, with the districts failing to achieve the 60 percent target in multiple grades and subjects (Becker et al., 2017). As a result, the pilot protocols were modified to require that calibration sessions and scoring sessions occur within school districts during the school year (National Center for the Improvement of Educational Assessment, 2020).

Learning Progressions

Another particularly prominent effort to implement elements of balanced assessment systems since *Knowing What Students Know* has been in the area of "learning progressions," sometimes also called "learning trajectories" (see Corcoran et al., 2009, for an early history; see Shepard, 2018, for a more recent one). Early advocates for learning progressions described their potential benefits as an alternative to traditional ways of thinking about standards, curriculum, assessment, and instruction (Corcoran et al., 2009). Through making and rigorously testing hypotheses about how children develop mastery in core concepts, learning progressions pair nicely with—and require—more balanced forms of assessment (National Research Council, 2007). That is, to assess student progress along a learning progression, one must employ an assessment system that is coherent, comprehensive, and continuous. Indeed, early conceptions of learning progressions emphasized their potential for promoting "[c]learer ties to instruction," "[providing] reference points for assessments that report in terms of levels of progress," and "[informing] the design of curricula that are efficiently aligned with what students need" (National Research Council, 2007, p. 9)—all clear nods to the principles of balanced assessment.

The concept of learning progressions has been influential, achieving a foothold in mathematics education (e.g., Clements & Sarama, 2020; Daro et al., 2011) and informing the design of the NRC's *A Framework for K–12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* (2012), which itself informed both the NGSS and other sets of state standards that have been adopted nearly nationwide (Duncan & Rivet, 2013). Shepard (2018) offers a thorough recent review of learning progressions, their development, and their impact. Shepard and colleagues (Shepard, 2018; Shepard et al., 2018) argue that learning progressions are best built "from the bottom up, focusing on local jurisdictions or curricular projects, where it is more likely to be possible to design for coherence among curriculum, instruction, assessment, and teacher learning" (Shepard, 2018, p. 167). Learning progressions have been built, primarily funded by the National Science Foundation, in domains such as matter and atomic-molecular theory, scientific argumentation, modern genetics, energy, evolution, and celestial motion. Shepard (2018) also notes the "extensive research and development and detailed materials" (p. 168) needed to deploy learning progressions, pointing to successful examples like the *Building Blocks* mathematics curriculum that embodies these principles.

Despite some promising small-scale findings, evidence of teaching and learning growth from learning progressions is modest. While there has been substantial funding

allocated toward learning progressions projects through the National Science Foundation, the impact at scale is limited. Most of the research described by Shepard and colleagues (2018) is small in scale, and besides influencing the development of recent standards, there is very little evidence of impact on teaching and learning beyond these controlled studies.

Smarter Balanced Assessment Systems

The Smarter Balanced Assessment was one of the two major Common Core–aligned testing consortia funded through the Obama administration's Race to the Top program in 2009 (The White House, n.d.). Although Common Core and its assessment consortia have become entangled in political battles across the United States since their initial rollout, as of 2023, 12 states are still Smarter Balanced members (Smarter Balanced, n.d.). While Smarter Balanced is best known for its state summative assessment, the organization also offers elements of its overall design to facilitate better balance in state and local assessment systems. For instance, Smarter Balanced offers partner states optional interim assessments in the form of interim assessment blocks, or brief assessments focused on just a few assessment targets, as well as interim comprehensive assessments that are built to align with and provide scores on the standard Smarter Balanced scale (Hardoin et al., 2020).

These interim assessment systems contain elements that would seem to enhance the balance of local assessment systems as compared to more traditional commercial interim assessments, which often have limited or uncertain alignment with larger summative tests. For instance, the Smarter Balanced interim assessments include a variety of item types. While the constructed response items offered must be scored locally, Smarter Balanced offers training to teachers on scoring. The interim assessments provide results to teachers rapidly and in a form that can be instructionally useful for informing remediation and differentiation. Smarter Balanced also offers Tools for Teachers, aimed at aligning to interim assessments and differentiating instruction based on student performance. According to an independent evaluation of Smarter Balanced in California, evaluators found some evidence that these reforms are widely used and well received in local districts. Where educators found the Smarter Balanced interim assessments less useful was generally where the assessments were mandated in a fixed schedule that did not align well with local curriculum coverage and pacing (Hardoin et al., 2020).

Other Examples in Practice

There are many other examples of innovations in assessment over the 20-plus years since *Knowing What Students Know* that could be said to embody elements of balanced assessment systems. Conley (2018) provides an overview of some of these examples, including case studies of a few districts. Conley does not use the term balanced in his book, but the systems he describes often include principles of balance. For a fuller treatment, we encourage interested readers to read Chapter 6 of *The Promise and Practice of Next Generation Assessment*, but note that the description of implementation and impact of the examples he cites is relatively thin (Conley, 2018).

One type of system Conley (2018) cites is a commercial curriculum system leading to a specialized diploma. He cites several similar examples of such systems, including International Baccalaureate, Cambridge International, and the AP Capstone program. These programs emphasize tight coherence among curriculum and assessment, as well as a developmental approach to teaching and measuring student learning. These systems also include multiple forms of assessment that are both interim and summative.

Another type of balanced assessment system is represented by district–researcher collaboratives under the umbrella of the Assessment for Learning Project (ALP). Sponsored by the Hewlett Foundation, the ALP was focused on "deeper learning" and often funded district collaboratives (Conley, 2018). Again, the projects funded under the ALP emphasized principles related to balanced assessment systems, such as greater emphasis on the instructional utility of assessments, a focus on growth in student mastery, and a comprehensive assessment approach based in multiple measures of student performance. Conley (2018) also provides a case study of the Summit Public Schools, which has worked to better integrate curriculum and assessment and, in collaboration with technology providers, more carefully monitor student growth.

FACTORS THAT HAVE HINDERED THE GROWTH AND IMPACT OF BALANCED ASSESSMENT SYSTEMS

Although the vision contained in *Knowing What Students Know* was inspiring enough to spur engagement from a generation of scholars, readers would be forgiven for asking whether, this intellectual engagement notwithstanding, balanced assessment systems have ever really "caught on." Struck by the seeming disconnect between continual intellectual ferment and seemingly modest impact in practice, in this final section, we ask "What went wrong?"—or perhaps more accurately, "What continues to go wrong?" Balanced assessment systems are intuitively appealing, and the ideas underlying these systems seem as if they would be widely supported, both by experts in the assessment and policy fields and by practicing educators. But the accomplishments associated with balanced assessment systems to date are vanishingly modest. Why is it that balanced assessment systems have failed to take hold across states and districts? What needs to change, moving forward, to realize better results?

In this section, we introduce key factors we think have contributed to difficulties in achieving balanced assessment systems in practice, echoing and expanding on an earlier analysis by Marion and colleagues (2019a) and the brief discussion in Chapter 1 of this volume, "Reimagining Balanced Assessment Systems: An Introduction." Here, we survey both the positive and the negative—what we think is needed to implement a balanced assessment system as compared with what currently exists. This section focuses mainly on large-scale state assessments—the "elephants in the room"—because they, due to federal policy requirements, still drive assessment policy and practice nationwide, and because the extant literature overwhelmingly focuses on state tests. Our recommendations are intended to inform the discussions in the rest of the chapters in this volume, so that the appealing ideas underlying balanced assessment systems might begin to take root in state and district assessment systems. We grouped the challenges into two categories: (1) technical and (2) political and practical (see Table 2-1).

TABLE 2-1 Factors Challenging Growth of Balanced Assessment Systems

Technical Challenges	Political and Practical Challenges
• Measuring Multiple Complex Domains • Interpreting Information Across Grade Levels for Multiple Dimensions • Weighting Multiple Measures • Scoring Student Work • Implementing and Adapting Technology	• Poorly Designed Assessment and Curriculum Policies • Shifting Political Barriers • Challenges of Embedding Assessments in the Curriculum • Lack of Capacity Across Levels of the System • Instructional Reform in the Context of Loosely Coupled Systems

Technical Challenges

As we have hinted throughout this chapter, obtaining balance in assessment systems is technically difficult for a variety of reasons. Conley (2018) identifies some of the most important technical challenges in next generation assessment. He does not, like many researchers working in this space, use the term balance, although the principles he describes are similar to those in balanced assessment systems. We have not repeated his treatment of this topic here, but briefly summarize and elaborate on key technical issues that have challenged efforts to develop balanced and next generation assessments alike. Experts seeking to address these technical challenges are likely to have differing views or arrive at alternative conclusions of how to manage them, depending on the specific context. Given the complexity involved, it is unlikely that districts will be able to find off-the-shelf assessments that are sufficiently curriculum-embedded to create a balanced system. More likely, districts would need to retain and consult experts as they design an assessment system with the desired level of balance or attempt to adapt an off-the-shelf option.

Measuring Multiple Complex Domains

Typically, standardized assessments focus on one content domain and, to a large extent, one or two levels of cognitive complexity—mostly memorization and procedures (see, e.g., Polikoff et al., 2011). Because they aim to capture a more authentic picture of what students know and can do, balanced assessment systems must measure complex thinking skills, and creating these measures is simply more technically challenging than the more traditional item types. Consider the NGSS and their three-dimensional structure—disciplinary core knowledge, crosscutting concepts, and scientific practices. The method of constructing and scoring a set of items measuring all three of these dimensions is no small task.

Interpreting Information Across Grade Levels for Multiple Dimensions

Balanced assessment systems will typically seek to make more productive use of the longitudinal nature of assessment than traditional systems, including by using vertical scales that span grades. Vertical scales have their own well-established challenges (see, e.g., Briggs, 2013). These challenges are only further compounded by the greater ambition of balanced assessment systems to measure more complex domains. Put simply,

it is hard enough to create a vertical scale for mathematics skills, but what does it look like to create one for the ability to solve complex real world problems?

Weighting Multiple Measures

In balanced assessment systems with multiple measures, scores must be aggregated. This is especially true for assessment systems that are used, in whole or in part, in accountability systems. Conley (2018) discusses the example of California's CORE districts and their attempts to combine multiple measures through complex weighting. Setting weights in these systems is typically more art than science, and relies heavily on value judgments from the practitioners and policymakers who seek to use the data for various purposes.

Scoring Student Work

Balanced assessment systems must include more complex tasks to measure more complex skills, as well as to make the results more instructionally useful to teachers—although we note that many would argue that state summative assessments can never be useful to teachers because they cannot inform classroom instruction, full stop. But more complex tasks are dramatically more onerous to score than simple tasks, as well as to validly score across classrooms and schools and in ways that teachers can use the resulting information to diagnose and address student learning needs—regardless of whether those complex tasks are part of a local or state level assessment. Conley (2018) discusses some strategies that can help with the scoring burden. One strategy is for teachers to eschew some of the "busy work" they would usually assign and score, which would allow them to focus their energies on the admittedly larger charge of scoring more complex tasks. Another strategy is for teachers to involve students in scoring—for instance, students giving each other peer feedback before the final due date. In this case, teachers can provide high-level feedback to the class based on examining all the students' work, but only provide scores for individual students. Schools can also structure schedules to facilitate the grading of more complex work. Conley nods to the complexity of this work but argues that

> scorers can achieve high levels of agreement … when they are properly trained in the use of a scoring guide, … use criterion-based decision-making processes, and are well trained on exemplars so that they … are able to apply mental models of what they are looking for. (Conley, 2018, p. 157)

While this is undoubtedly the case, finding the time and resources to properly train teachers on scoring guides and exemplars, as well as establishing and maintaining high levels of agreement given the inevitable turnover in personnel, poses another obstacle for districts and schools seeking to implement balanced assessments.

Implementing and Adapting Technology

Finally, Conley (2018) discusses the opportunities and limitations of technology in the context of next-generation assessment. Certainly, technology can offer advantages

in terms of administering and scoring more complex item types. Technology can also aid in producing scores more rapidly and presenting student performance in ways that are more usable to teachers. However, both teachers and students need to able to use the technology in order for these opportunities to materialize, and that is far from a given. Additionally, some of the best technological innovations will come from smaller startup education technology companies, but Conley (2018) notes that these companies are often disadvantaged in procurement processes.

Political and Practical Challenges

Even if these technical challenges could be overcome at a large scale, there are also important political and practical barriers that impede more balanced assessment systems from taking root in districts and states.

Poorly Designed Assessment and Curriculum Policies

Balanced assessment systems are complicated to enact, and without incentives and support it is unlikely that implementation will involve anything more than isolated instances of local implementation. Unfortunately, there have been few incentives to develop assessment systems in line with *Knowing What Students Know.* Although it is worth noting that the New Hampshire pilot discussed above was facilitated by the U.S. Department of Education's NCLB waiver process and now the Every Student Succeeds Act (ESSA) authorized, Innovative Assessment Demonstration Authority. Far from encouraging the implementation of balanced assessment systems, there are a variety of ways in which state and federal assessment and curriculum policies currently undermine their spread.

As an example, consider the ways that federal assessment peer review guidance could support balanced assessment systems but instead falls short. Historically, peer review guidance has required that states base accountability decisions on a single summative year-end assessment of student knowledge and skills related to grade-level standards. This requirement makes sense from the standpoint of generating a point-in-time estimate of student proficiency against grade-level standards, but it conflicts directly with the utility of the test results for improving teaching and learning. At best, the results of current accountability assessments could be useful for teachers in the next academic year, but even a passing conversation with practicing educators makes clear that state accountability tests are generally viewed as virtually useless for informing instructional choices. Indeed, many would argue that this is explicitly not the purpose of these assessments, although the messaging around the intended uses of state summative tests is far from clear. This type of requirement also directly contradicts the principles of balanced assessment systems, in that assessments with stakes should be based on a range of types of evidence (i.e., comprehensiveness).

Federal testing and accountability policy under NCLB emphasized the importance of "percent proficient" as the primary metric for school effectiveness. This approach takes the wealth of assessment data available for each test taker, boils it down to a single score, and then dichotomizes that score to either above or below proficient. Under NCLB, growth-based approaches to assessing students or evaluating school

effectiveness were verboten. More recently, under ESSA, federal policy has allowed for growth-based measures of performance but still emphasizes that states must place a heavy emphasis on grade-level proficiency (Every Student Succeeds Act, 2015). These requirements directly contradict the balanced assessment criterion of continuity.

The third pillar of balanced assessments is coherence, and here too federal assessment policy hinders adoption or expansion. Federal assessment policy says nothing of consequence about curriculum, as federal policy intentionally stays far from curriculum issues, and the result is that states vary considerably in their effort—or lack thereof—to ensure students have access to aligned curriculum materials and tightly coupled assessments. Federal policy also does not give so much as a passing nod to theories or models of learning. Theories of learning are not typically emphasized in state standards, although, as is mentioned previously, the NGSS were informed by learning progressions in science.

ESSA is widely seen to offer more opportunities than NCLB for states to improve their assessment systems, moving them more in line with the principles of balanced assessment systems. For instance, ESSA requires states to extend beyond only reading and math proficiency, allows states to use growth-based measures of achievement, and permits more innovative forms of assessment (Conley, 2018). Still, federal requirements and regulations, such as the requirement that every student be tested and receive an individual score indicating their mastery of grade-level content, have substantial impact on states' decisions about the design and implementation of their assessment systems. The result is, despite the modest affordances of ESSA, state assessment systems look like the systems that have been historically required under federal law, not the kinds of systems advocates of balanced assessment systems would prefer.

This argument is not to say that federal policy could not support balanced assessment system principles in practice—quite the contrary. Policy tools exist to encourage or require states to adopt better assessments, but these policy tools are not being used. The most straightforward tool—and the one that the federal government has been most adept at using—is money. But there are other tools as well, including clear and specific guidance, regulation, and enforcement aligned with balanced assessment system principles. For instance, federal policies could encourage innovation in assessment systems, set high bars for the use of assessments for consequential decisions, and encourage states to facilitate tight assessment and curriculum alignment. We return to some of these issues in more detail below, and Chapter 9 of this volume, "Policy Influences on Ambitious Classroom Instruction, Assessment, and Learning," includes more thoughts on the ways policy can influence instruction and assessment.

Shifting Political Barriers

To create and sustain complex educational reforms like balanced assessment systems, substantial political and structural challenges must be overcome. The political barriers to educational reform have been well described elsewhere (e.g., Polikoff, 2021), but are worth briefly elaborating on here as well. First, and perhaps the defining characteristic of American education, is its decentralization—it includes 50 states and 13,000 school districts, each with their own elected and appointed officials, and each creating and seeking to implement policy (Polikoff, 2021). Without some level of top-

down leadership on assessment issues, wide-scale adoption of balanced assessment system features is likely impossible. However, there is often profound resistance from the public to the perception that states or the federal government are usurping local authority—a seemingly permanent source of tension.

At each level of the decentralized American educational system, political leadership is often unstable, with rapid fluctuations from party to party—and even within a party over time as priorities and goals change. Consider, for instance, the rapid shifts in guidance related to transgender students and Title IX requirements as federal administrations changed during the 2010s (Hersher & Johnson, 2017). Although assessment policy may be somewhat insulated from this instability—it has endured over multiple decades across both Republican and Democratic administrations—it is likely that more ambitious assessment reforms would run the risk of falling victim to political instability of one form or another as happened to the Common Core assessment consortia (Jochim & McGuinn, 2016).

But beyond the mere instability itself, there are the challenges associated with elected or appointed educational leadership positions. The disconnect between rhetorical cycles of educational reform and the time necessary to secure real change is an old problem (Tyack & Cuban, 1997). More often than not, the desire among elected officials is to have short-term political victories, which are usually characterized by a claimed improvement in some type of student outcomes (Tyack & Cuban, 1997). There seems to be little appetite for the sustained, hard work that would be required to build and maintain complex policy instruments like balanced assessment systems. These systems require both infusions of initial capital and sustained resources over time, and are unlikely to produce the short-term bumps in performance that many elected officials would like to be able to point to.

Of course, there are counterexamples that show what is possible with sustained vision and leadership focused on appropriately using available levers of government. For example, Louisiana's reforms started under State Superintendent John White but have continued for more than 10 years and have substantially revised the state's approach to curriculum and assessment (Kaufman et al., 2016, 2018). Louisiana built its curriculum reforms around a coherent theory of change, aligning key elements like professional learning around their curriculum-driven vision. The state created powerful incentives for local school districts to adopt and use high-quality curriculum materials, using the power of the Louisiana Department of Education to rapidly encourage adoption. It also provided or identified providers of aligned professional development, leading to sustained teacher learning. When Secretary White stepped down, these reforms had become embedded in Louisiana's educational culture, and persisted into the subsequent administration. By creating a coherent vision and building a supportive constituency through careful policy design, this approach ensured the longevity of the reforms (Kaufman et al., 2016, 2018).

Challenges of Embedding Assessments in the Curriculum

Curriculum-embedded assessments are at the heart of balanced assessment systems. A balanced assessment system requires tight linkages among assessment, curriculum, and instruction; and, ideally, assessment systems will not merely be aligned

with instruction and curriculum, but seamlessly integrated into instruction. While the goal of curriculum-embedded assessments is admirable, this goal is currently far from the curriculum and instructional reality of American schools.

In most U.S. states, and in many grades and subjects even within the most curriculum-active states, there is very little curriculum centralization (Polikoff, 2021). Many states have no guidance about what curriculum materials schools and districts should adopt. Other states put out lists of approved materials in some subjects and grades, but make those lists strictly advisory. Modest incentives or requirements for districts and schools to adopt particular materials are only offered in a vanishingly small number of state, subject, and grade combinations (Polikoff, 2021). Almost no states keep track of which materials are being used where—and even when they do, the information is often unreliable (Hutt & Polikoff, 2020). What little data we have suggests that there is very little consistency across districts in which materials they adopt (Polikoff, 2021). This fact alone is almost fatal to the idea of state-driven, curriculum-embedded balanced assessment systems—without greater centralization in curriculum decisions, it is hard to see how states can meaningfully support curriculum-embedded balanced assessment systems. To be sure, local actors could still build higher-quality and embedded assessments in their own adopted materials, but this would require substantial capacity—we discuss this possibility below.

Beyond formally adopted curriculum materials, there is the question of how teachers make use of the curriculum materials they are given. Again, the reality of the American educational system is that teachers typically use core curriculum materials as one source among many for instructional guidance. Teachers in U.S. classrooms overwhelmingly engage in various forms of curriculum supplementation (Silver, 2022). Survey data indicate that nearly all teachers supplement with materials from the Internet, from their own repositories, and with materials they create, often with staggering frequency. Curriculum and instruction are indeed the single domain over which individual teachers have the most control (Ingersoll, 2006). Again, the extent of teacher authority over curriculum—and the degree to which teachers exercise that authority by modifying, adding to, or subtracting from the formally adopted curriculum—is something of a stake in the heart of the idea of widespread adoption and implementation of balanced assessment systems.

These realities about curriculum control in U.S. schools and classrooms run headlong into the goal of widespread curriculum-embedded assessments. One path through these challenges is a more assertive state role in curriculum decisions, something that has been advocated and discussed at length elsewhere (Polikoff, 2021). Briefly, this path would include stronger state guidance or requirements for the selection of curriculum from among a small set of high-quality options, coupled with the creation and use of embedded assessments in those same materials. If all districts in a state were using, for instance, one of three highly regarded curriculum materials—and if the state, or the curriculum provider itself, could create and support embedded, ongoing assessment—this could offer a path toward balance. Louisiana's recent waiver from the U.S. Department of Education to develop a curriculum-embedded assessment system shows what is possible when a state has greater centralization and control over curriculum, although that effort appears too early to have had meaningful evaluation (NWEA, 2021). Louisiana's approach addresses several of the principles of balanced assessment—most notably, it

is curriculum-embedded (which can only work in a state where substantial proportions of districts use the same materials), drawing on the English language arts (ELA) texts and content that students have used throughout the school year.

But even Louisiana's approach will run up against the realities of how U.S. teachers use core curriculum materials. The scope and nature of curriculum supplementation is an issue that policy barely attempts to address, but that must be tackled in order for a balanced assessment system to take root. Certainly, there is little interest in meaningfully restricting teachers' curriculum control, but there may be ways to productively redirect curriculum supplementation in ways that support, rather than undermine, the core curriculum and its embedded assessments. For instance, building collaborative structures and clear expectations that encourage teachers to collaboratively supplement within schools and districts could allow for sufficient between-classroom consistency that would allow balanced assessment systems to become more locally feasible (see Polikoff, 2021, for more discussion of this vision).

Lack of Capacity Across Levels of the System

Over the past two decades of standards-based assessment, states and districts have developed substantial experience in implementing assessments, using assessment data, and messaging assessment results to families and other stakeholders. But balanced assessment systems are much more complex than the traditional assessment systems they seek to replace. For instance, they require multiple measures, not just one, to make important decisions. They require greater timeliness in reporting—and, simultaneously, more sophisticated forms of evidence from more complex items. They necessitate deeper, shared understanding among educators across classrooms and grade levels, as well as more seamless integration of assessments and their results in the curriculum. In short, they require greater capacity for designing, carrying out, and using assessments across actors in the system.

The implication of implementing balanced assessment systems is that there needs to be substantial assessment capacity in the nation's educational systems, and if that capacity does not already exist, that effective and ongoing capacity building will take place. However, assessment literacy has always been a sore spot for our educational systems (Popham, 2009). Teacher education programs have historically spent little, if any, time covering assessment literacy (Stiggins, 2006), and there is little evidence to suggest this has changed (Popham, 2018). After *Knowing What Students Know*, researchers like Stiggins (2006) laid out principles for teacher in-service and pre-service education in order to build teachers' assessment literacy, but these changes to existing protocol have not happened. Without a substantial increase in the assessment capacity of individual educators, achieving the vision of classroom-driven balanced assessment systems advocated in this volume is likely impossible. Chapter 5 of this volume, "Assessment Literacy and Professional Learning," offers some thoughts on how professional learning can support balanced assessment systems.

There are many reasons for the failure to build assessment literacy across the system, and these are correlated with the issues already discussed in this chapter. Teacher education, both pre-service and in-service, is highly decentralized, with thousands of teacher training programs in operation and very little in the way of standardized expectations.

Teacher educators themselves are often highly resistant to assessment-driven reform (Cochran-Smith, 2006), although they might be more receptive to balanced assessment systems than more traditional forms of test-driven accountability. Finally, in-service teacher learning opportunities are notoriously poor in both design and impact (Darling-Hammond et al., 2017). These are all substantial barriers to overcome.

Perhaps due to the difficulty in achieving assessment literacy through policy, commercial providers have stepped in. For example, large-scale interim assessment providers like the Northwest Evaluation Association and Curriculum Associates provide assessments to thousands of school districts. These assessments can be curriculum-embedded (e.g., in the case of Curriculum Associates, which offers a companion curriculum), but are not necessarily so. The assessments can also be "embedded" in the more vernacular sense of the term, in that they are scheduled to occur at fixed time points during the school year, but not meaningfully embedded in terms of the content they emphasize. Indeed, claims of alignment of interim assessments with curriculum or standards regularly go unverified (Perie et al., 2007). These assessments have met a need that districts had for reasonably high-quality assessments that could be quickly analyzed and used to provide feedback on student progress throughout the year, but they often have fallen far short of contributing to balance in practice.

Instructional Reform in the Context of Loosely Coupled Systems

The loose coupling (Weick, 1976) that characterizes educational systems in the United States makes complex reform extremely challenging (e.g., Labaree, 2012). Key elements of loosely coupled systems include an absence of regulation, the failure of leaders to influence subordinates, decentralization of power, autonomy of ground-level employees, and a lack of consensus around goals (Weick, 1976). While these characteristics thwart substantial reform efforts, they also can serve advantageous or protective functions, such as allowing the organizations to endure constantly changing environments, permitting failures in some systems without damaging the broader organization, and enabling local adaptation (Labaree, 2012).

Loose coupling has contributed to the standards movement's lack of success in the last several decades (Polikoff, 2021). Regarding standards-based reforms, states have largely left difficult implementation decisions to local policymakers (e.g., decisions around teacher learning and curriculum adoption). As a result, teachers have almost never received the types of clear guidance needed to understand, let alone implement, complex instructional policies. These challenges have become even more fraught with increasingly complex college- and career-ready standards (Polikoff et al., 2022), which move topics across grades, include more emphasis on conceptual understanding, and often include additional dimensions like mathematics or science practices on top of content expectations. These challenges create an inertia for existing practices that is difficult to overcome.

One way to understand education reform since the 1990s is as various efforts to try and more tightly couple levels of the system, including federal policy to state policy, state policy to student learning outcomes, and state policy to teacher instruction. These efforts have been limited by the factors outlined in this chapter, including shifting politics and policies, the limited capacity in the system, and the increasing ambitions for

our standards and assessment systems (e.g., Cohen et al., 2022). Balanced assessment systems, too, represent a highly ambitious reform at the intersection of assessment and instruction. This chapter's analysis points to the need to commit to ongoing development of assessment systems while simultaneously working to create tighter couplings in order to see more meaningful implementation.

CONCLUSION

Who could be opposed to a balanced assessment system? Certainly no one wants imbalance. Yet, most would agree that our assessment systems are currently and have been imbalanced. They were imbalanced when *Knowing What Students Know* was first published, and they are imbalanced today, although perhaps in different ways. The role and quality of state summative tests has ebbed and flowed over time, while the use of interim assessments has exploded since the publication of *Knowing What Students Know*. At the same time, the technology to bring about balance has grown. Advances in assessment quality, spurred in part by the Common Core, have brought better large-scale assessments (Doorey & Polikoff, 2016). Some states have also increased their authority over curriculum materials, making it far more possible for curriculum-embedded assessments to take hold than in a laissez-faire curriculum market.

Still, the nation is far from achieving balance in assessment systems at scale, and the purpose of this chapter is to discuss some reasons for this failure. In terms of lessons learned, we think there are several.

First, achieving balance must be made both more understandable and feasible for educators and local and state policymakers. The criteria underlying balanced assessment systems are laudable, but the ideas are too complex for widespread comprehension and implementation in the current highly decentralized, capacity-poor education systems. Also, even when there is general agreement on the underlying principles, the proliferation of similar ideas with different terminology has added confusion and created the sense that even similar-minded districts are pursuing different paths. It is likely that state departments of education, perhaps working in concert with curriculum developers and providers, must play a larger role in giving local actors clear guidance on how to make assessment systems more balanced. It cannot be "here is the guidance, go forth and conquer;" it must be closer to "here is what you should do, and here are some tools you can use to do it." Despite the need for this clear guidance, as we have noted throughout, we do not think the literature is clear and specific enough in describing examples of balanced assessment systems and demonstrating their efficacy. One further challenge is that states may lack the necessary capacity—either technical or political—or the will to offer this extra level of support. But acknowledging this problem only underscores the point that this work must be centralized, as these difficulties are only compounded when left to individual districts.

Second, a national policy discussion about the role of state summative assessments in accountability is needed. The status quo presents a situation in which state standardized tests are limping along, supported weakly by many but strongly by few; accountability uses have diminished but educators still feel considerable pressure to tend to their students' test scores; and state tests are widely pilloried for not providing useful data to inform instruction—a purpose they were never well suited to serve. This

situation serves no constituency well. To do better, an approach to large-scale assessment that ensures appropriate safeguards to protect the rights of underserved students while minimizing the distorting consequences on teaching and learning is needed. Such an approach could also make space for higher-quality and more useful curriculum-embedded assessments that would improve overall balance. Without changes in the state and federal policy context around assessment, balance will be unachievable at lower levels, and the proverbial cart will continue to drive the horse.

Third, assessment experts must be more honest and realistic about the utility of better assessment systems. There is a long history of assessment innovations being oversold, and balanced assessment systems are no different. We believe that well-designed efforts to bring about greater assessment balance would be beneficial, but they would, like all education policy innovations, provide an incremental improvement, not a revolutionary one. They must be coupled with other policies known or strongly suspected to improve student learning, including more generous and equitable school financing; high-quality, highly usable curriculum materials; more and better use of instructional time; and more well-trained educators.

The two decades since the publication of *Knowing What Students Know* have provided ample time for the field to relearn a very old lesson in school reform: describing a better way to do things is never enough to bring about change. Only by tending to the political and organizational demands of reforming ideas can we ever hope to secure a place for them in our schools.

REFERENCES

Becker, D. E., Thacker, A. A., Sinclair, A., Dickinson, E. R., Woods, A., & Wiley, C. R. H. (2017). *Formative evaluation of New Hampshire's Performance Assessment of Competency Education (PACE)*. HumRRO. https://www2.ed.gov/policy/elsec/guid/stateletters/nhpaceformativevalrpt2017.pdf.

Briggs, D. C. (2013). Measuring growth with vertical scales. *Journal of Educational Measurement*, *50*(2), 204–226.

Brookhart, S. M. (2013). Comprehensive assessment systems in service of learning: Getting the balance right. In R. W. Lissitz (Ed.), *Informing the practice of teaching using formative and interim assessment: A systems approach* (pp. 165–184). Information Age Publishing.

Bush, G. W. (2000). *Speech to the NAACP's 91st Annual Convention*. https://www.washingtonpost.com/wp-srv/onpolitics/elections/bushtext071000.htm.

Clements, D. H., & Sarama, J. (2020). *Learning and teaching early math: The learning trajectories approach* (3rd ed.). Taylor and Francis.

Cochran-Smith, M. (Ed.). (2006). *Policy, practice, and politics in teacher education*. Corwin Press.

Cohen, J., Hutt, E., Berlin, R., & Wiseman, E. (2022). The change we cannot see: Instructional quality and classroom observation in the era of Common Core. *Educational Policy*, *36*(6), 1261–1287. https://doi.org/10.1177/0895904820951114.

Conley, D. T. (2018). *The promise and practice of next generation assessment.* Harvard Education Press.

Corcoran, T., Mosher, F. A., & Rogat, A. (2009). *Learning progressions in science: An evidence-informed approach to reform*. Center on Continuous Instructional Improvement, Teachers College, Columbia University. https://files.eric.ed.gov/fulltext/ED506730.pdf.

Darling-Hammond, L., Hyler, M. E., & Gardner, M. (2017). *Effective teacher professional development.* Learning Policy Institute.

Daro, P. Mosher, F. A., & Corcoran, T. (2011). *Learning trajectories in mathematics: A foundation for standards, curriculum, assessment, and instruction.* Consortium for Policy Research in Education. https://files.eric.ed.gov/fulltext/ED519792.pdf.

Doorey, N., & Polikoff, M. (2016). *Evaluating the content and quality of next generation assessments*. Thomas B. Fordham Institute. https://files.eric.ed.gov/fulltext/ED565742.pdf.

Duncan, R. G., & Rivet, A. E. (2013). Science learning progressions. *Science*, *339*(6118), 396–397. https://doi.org/10.1126/science.1228692.

Evans, C. M., & Lyons, S. (2017). Comparability in balanced assessment systems for state accountability. *Educational Measurement: Issues and Practice*, *36*(3), 24–34. https://doi.org/10.1111/emip.12152.

Every Student Succeeds Act. 20 USC § 6301. (2015). https://www.congress.gov/114/plaws/publ95/PLAW-114publ95.pdf.

Goals 2000: Educate America Act. 20 USC § 5801. (1994). https://www.congress.gov/103/statute/STATUTE-108/STATUTE-108-Pg125.pdf.

Hardoin, M. M., Dvorak, R. N., Dickinson, E., Paulsen, J., Gribben, M., & Revivo, R. (2020). *California Assessment of Student Performance and Progress (CAASPP): 2020 independent evaluation report (Volume 1)*. HumRRO. https://www.cde.ca.gov/ta/tg/ca/documents/caaspp20evalrptvol1.pdf.

Harris, K. (2020). Forty years of falling manufacturing employment. *Beyond the Numbers: Employment and Unemployment*, *9*(16).

Hersher, R., & Johnson, C. (2017, February 22). Trump administration rescinds Obama rule on transgender students' bathroom use. *NPR The Two-Way*. https://www.npr.org/sections/thetwo-way/2017/02/22/516664633/trump-administration-rescinds-obama-rule-on-transgender-students-bathroom-use.

Hutt, E., & Polikoff, M. S. (2020). Toward a framework for public accountability in education reform. *Educational Researcher*, *49*(7), 503–511. https://doi.org/10.3102/0013189X20931246.

Improving America's Schools Act of 1994. 20 USC § 2701. (1994). https://www.govtrack.us/congress/bills/103/hr6/text.

Ingersoll, R. M. (2006). *Who controls teachers' work?: Power and accountability in America's schools*. Harvard University Press.

Jochim, A., & McGuinn, P. (2016). The politics of the Common Core assessments: Why states are quitting the PARCC and Smarter Balanced testing consortia. *Education Next*, *16*(4), 44–53.

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

Kaufman, J. H., Cannon, J. S., Culbertson, S., Hannan, M. Q., Hamilton, L. S., & Meyers, S. (2018). *Raising the bar: Louisiana's strategies for improving student outcomes*. RAND Corporation. https://www.rand.org/pubs/research_reports/RR2303.html.

Kaufman, J. H., Thompson, L. E., & Opfer, V. D. (2016). *Creating a coherent system to support instruction aligned with state standards: Promising practices of the Louisiana Department of Education*. RAND Corporation. https://www.rand.org/pubs/research_reports/RR1613.html.

Labaree, D. F. (2012). *Someone has to fail: The zero-sum game of public schooling*. Harvard University Press.

Lyons, S., Evans, C., Marion, S., & Thompson, J. (2017). *New Hampshire Performance Assessment of Competency Education technical manual*. National Center for the Improvement of Educational Assessment.

Marion, S., Thompson, J., Evans, C., Martineau, J., & Dadey, N. (2019a). *The challenges and opportunities of balanced systems of assessment: A policy brief*. National Center for the Improvement of Educational Assessment. https://files.eric.ed.gov/fulltext/ED598421.pdf.

Marion, S., Thompson, J., Evans, C., Martineau, J., & Dadey, N. (2019b). *A tricky balance: The challenges and opportunities of balanced systems of assessments*. Paper presentation at National Council on Measurement in Education, 2019, Toronto, Ontario.

McGuinn, P. J. (2006). *No Child Left Behind and the transformation of federal education policy, 1965–2005*. University Press of Kansas.

National Center for the Improvement of Educational Assessment. (2020). *New Hampshire's innovative assessment system: Performance Assessment of Competency Education (PACE): Evaluating technical quality (Volume 2): Results*. https://www.education.nh.gov/sites/g/files/ehbemt326/files/files/inline-documents/pacemanualvol2results.pdf.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Government Printing Office. http://edreform.com/wp-content/uploads/2013/02/A_Nation_At_Risk_1983.pdf.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics.* National Council of Teachers of Mathematics. https://archive.org/details/curriculumevalua00nati/mode/2up.

National Council on Education Standards and Testing. (1992). *Raising standards for American education*. Government Printing Office. https://files.eric.ed.gov/fulltext/ED338721.pdf.

National Research Council. (1999). *How people learn: Bridging research and practice*. National Academy Press. https://doi.org/10.17226/9457.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. National Academy Press. https://doi.org/10.17226/10019.

National Research Council. (2003). *Assessment in support of instruction and learning: Bridging the gap between large-scale and classroom assessment: Workshop report.* The National Academies Press. https://doi.org/10.17226/10802.

National Research Council. (2006). *Systems for state science assessment*. The National Academies Press. https://doi.org/10.17226/11312.

National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K–8*. The National Academies Press. https://doi.org/10.17226/11625.

National Research Council. (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas.* The National Academies Press. https://doi.org/10.17226/13165.

National Research Council. (2014). *Developing assessments for the Next Generation Science Standards*. The National Academies Press. https://doi.org/10.17226/18409.

New Hampshire Department of Education. (2023). *Performance Assessment of Competency Education*. https://www.education.nh.gov/who-we-are/division-of-learner-support/bureau-of-instructional-support/performance-assessment-for-competency-education.

No Child Left Behind Act of 2001. 20 USC § 6301. (2001). https://www.congress.gov/bill/107th-congress/house-bill/1/text.

NWEA. (2021). *Louisiana and NWEA: Creating innovative assessments to foster equity and deeper learning*. NWEA State Solutions. https://www.nwea.org/resource-center/white-paper/47609/Louisiana-IADA-Brief_NWEA_whitepaper-1.pdf.

Perie, M., Marion, S., Gong, B., & Wurtzel, J. (2007). *The role of interim assessments in a comprehensive assessment system*. Aspen Institute.

Polikoff, M. (2021). *Beyond standards: The fragmentation of education governance and the promise of curriculum reform*. Harvard Education Press.

Polikoff, M. S., Desimone, L. M., Porter, A. C., Garet, M. S., Stornaiuolo, A., Pak, K., Flores, N., Smith, T. M., Song, M., Fuchs, L. S., Fuchs, D., & Nichols, T. P. (2022). *The enduring struggle of standards-based reform: Lessons from a national research center on college and career-ready standards.* EdWorkingPaper 22-622. Annenberg Institute at Brown University.

Polikoff, M. S., Porter, A. C., & Smithson, J. (2011). How well aligned are state assessments of student achievement with state content standards? *American Educational Research Journal*, *48*(4), 965–995. https://doi.org/10.3102/0002831211410684.

Popham, W. J. (2018). *Assessment literacy for teachers in a hurry*. ASCD.

Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory Into Practice*, *48*(1), 4–11. https://doi.org/10.1080/00405840802577536.

Resnick, D. P. (1980). Minimum competency testing historically considered. *Review of Research in Education*, *8*, 3–29. https://doi.org/10.2307/1167122.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction*. Kluwer Academic Publishers. https://doi.org/10.1007/978-94-011-2968-8_3.

Shepard, L. A. (2018). Learning progressions as tools for assessment and learning. *Applied Measurement in Education*, *31*(2), 165–174. https://doi.org/10.1080/08957347.2017.1408628.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, *29*(7), 4–14. https://doi.org/10.3102/0013189X02900700.

Shepard, L. A., Penuel, W. R., & Pellegrino, J. W. (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. https://doi.org/10.1111/emip.12189.

Silver, D. (2022). A theoretical framework for studying teachers' curriculum supplementation. *Review of Educational Research*, *92*(3), 455–489. https://doi.org/10.3102/00346543211063930.
Smarter Balanced. (n.d.). The power of partnerships and collaboration. https://smarterbalanced.org/our-vision/partnerships.
Stiggins, R. J. (2006). Assessment for learning: A key to student motivation and learning. *EDge*, *2*(2), 1–19.
Stiggins, R. J. (2001). The unfulfilled promise of classroom assessment. *Educational Measurement: Issues and Practice*, *20*(3), 5–15. https://doi.org/10.1111/j.1745-3992.2001.tb00065.x.
The White House. (n.d.). Race to the Top. https://obamawhitehouse.archives.gov/issues/education/k-12/race-to-the-top.
Tyack, D., & Cuban, L. (1997). *Tinkering toward utopia: A century of public school reform*. Harvard University Press.
Tröhler, D. (2014). Change management in the governance of schooling: The rise of experts, planners, and statistics in the early OECD. *Teachers College Record*, *116*(9), 1–26. https://doi.org/10.1177/016146811411600903.
Weick, K. E. (1976). Educational organizations as loosely coupled systems. *Administrative Science Quarterly*, *21*(1), 1–19.