Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
Spring 2018, Vol. 37, No. 1, pp. 21–34

Using Learning and Motivation Theories to Coherently Link Formative Assessment, Grading Practices, and Large-Scale Assessment

L. A. Shepard and W. R. Penuel, *University of Colorado Boulder,* and J. W. Pellegrino, *University of Illinois at Chicago*

To support equitable and ambitious teaching practices, classroom assessment design must be grounded in a research-based theory of learning. Compared to other theories, sociocultural theory offers a more powerful, integrative account of how motivational aspects of learning—such as self-regulation, self-efficacy, sense of belonging, and identity—are completely entwined with cognitive development. Instead of centering assessment within systems that support use of interim and end-of-year standardized tests, we argue for a vision of formative assessment based on discipline-specific tasks and questions that can provide qualitative insights about student experience and thinking, including their identification with disciplinary practices. At the same time, to be consistent with a productive formative assessment culture, grading policies should avoid using points and grades "to motivate" students but should create opportunities for students to use feedback to improve their work. We argue for districts as the locus for the design of such coherent curriculum, instruction, and assessment activity systems because districts have responsibility for curriculum, teacher professional development, and equity; and districts allocate resources for textbooks and assessment.

Keywords: classroom assessment, formative assessment, learning theory, grading

Measurement researchers have recently expressed an interest in classroom assessment intended to support student learning. Growing interest in diagnostic classification models and learning progressions, for example, represents a significant departure from the field's nearly exclusive focus on psychometric models and test development methods needed for large-scale assessment. Measurement science could make meaningful contributions to classroom formative assessment and grading practices, especially regarding the definition of constructs and valid representation of intended learning goals. But there is also the danger that measurement frameworks could distort high-quality instruction, if the emphasis was on quantification rather than the qualities of student thinking or if they encouraged "testing" formats—necessary for the efficiencies of large-scale assessment—that are then replicated in computerized testing systems that track students' mastery of standards. Such testing formats are unnecessary in classroom contexts but often come to dominate classroom assessment.

This article is addressed to researchers in the measurement community and to educators responsible for making curriculum and assessment decisions for their classroom, school, district, or state. We explain why assessment design and validation must be grounded in an adequate, research-based theory of learning (National Research Council, 2001) if assessment is to be a support rather than a hindrance to quality teaching and learning that is accessible to all students. We also examine how implicit and explicit theories of learning shape the design of assessments for different purposes and at different levels of the educational system. We consider what this looks like when classroom and district level theories are conceptually compatible or *vertically coherent* as envisioned in *Knowing What Students Know* (National Research Council, 2001), and we acknowledge what might be done when districts or states lack curricular authority to enable such coherence.

In the first half of the article, focused on learning theory, our argument is laid out as follows. First, we explain a bit more about "models of cognition" and what it means to use these "theories" or models as the basis for assessment design. We provide a brief section on "big and little theories" identifying both broad, general theories regarding the nature of learning and development and particularized, fine-grained, discipline-specific models of learning in practice. We summarize the case made previously by Penuel and Shepard (2016a, b) as to why sociocognitive and sociocultural learning theories are more likely than predecessor theories to support ambitious teaching practices and to further equity. In the last

L. A. Shepard, Distinguished Professor and Dean Emerita, Research and Evaluation Methodology, School of Education, Room 235, University of Colorado Boulder, Boulder, CO 80309; lorrie.shepard@ucolorado.edu. W. R. Penuel, Professor, Learning Sciences and Human Development, The BUENO Center, Room 320-E, University of Colorado Boulder, Boulder, CO 80309; william.penuel@colorado.edu. J. W. Pellegrino, Distinguished Professor of Education, Learning Sciences Research Institute, University of Illinois at Chicago, 1240 West Harrison Street, Suite 1535, Chicago, IL 60607; pellegjw@uic.edu.

learning-theory section, we explain why discipline-specific, detailed models of learning are essential to the co-design of curriculum, instruction, and assessment.

In the second half of the article, we turn to considerations well known to measurement specialists about how assessment designs must be tailored to their specific purposes. We take up assessment purposes that occur the most often at different levels of the education system, which, if designed well, hold the greatest promise for designing assessment systems that are coherent and synergistic across levels. Thus, we address (1) formative assessment in classrooms as part of ongoing instruction, (2) classroom summative grading practices, (3) district-level curriculum-based program evaluation, and (4) state-level standards-based assessment and accountability. Although there are a number of design features that vary in response to demand characteristics for each of the levels and purposes—for example, timing, frequency, immediacy of feedback, need for standardization, and so forth—we focus especially on curricular specificity, because curriculum materials, embedded assessments, and accompanying professional development are the means by which theories of learning come to be enacted in classrooms and potentially could be made coherent across levels of the system.

In the concluding section, we acknowledge that drawing from both learning theory and measurement research to build equitable and coherent systems of assessment is a challenging task. But, we argue that it is also doable. There are existing proofs that such endeavors are possible, through research-practice partnerships and when districts—with authority to do so—are able to direct curriculum, professional development, and testing resources to this integrated purpose.

Learning Theory as the Basis for Assessment Design and Use

The *Knowing What Students Know* (National Research Council, 2001) report had special importance for the field of educational measurement because it was produced by an expert panel expressly convened by the National Academy of Sciences "to review and synthesize advances in the cognitive sciences and measurement and to explore their implications for improving educational assessment. At the heart of the committee's work was the critical importance of developing new kinds of educational assessments that better serve the goal of equity" (p. 1). *Knowing What Students Know* put forth the core idea that "models of cognition and learning" should provide "a basis for the design and implementation of theory-driven instructional and assessment practices" (p. 5). This was not merely a wishful recommendation but, in fact, reflected mounting evidence from decades of research that provided empirical support for different routes to learning in classrooms, particularly in mathematics and science education. The cognition "corner" of an "assessment triangle" (p. 44) provides the theory or set of beliefs by which observation methods are then designed to elicit evidence and interpretive frameworks are used to draw inferences about students' competencies.

Basing assessment on a model or theory of learning seems similar to the old idea in measurement and in teacher education of building assessments based on instructional goals and objectives, but it is actually a more challenging idea. Models of cognition are *research-based* when they are derived from studies of how students come to have the desired skills and abilities in domains of instruction, and they are *developmental*, meaning that the intermediate steps to full proficiency are examined and documented as part of the model, along with the means of support for helping students achieve proficiency. Thus, assessment frameworks or domain specifications must be conceptualized longitudinally, rather than cross-sectionally. This might mean documenting what moves learning forward within a single lesson or sequentially over longer term trajectories.

Integrating research on learning and knowing into the process of assessment design speaks directly to the goal of using assessment to improve teaching and learning. More specifically, *Knowing What Students Know* identified the underlying model of learning as the key means for ensuring that assessment systems function coherently and effectively. *Horizontal coherence* refers to the mutual support and compatibility that occurs when curriculum, instruction, and assessment are co-designed based on the same theory of learning. At the classroom level, this means that assessment occasions are integrated with instructional activities, though often unmarked for students as instances of evaluation. As mentioned previously, *vertical coherence* occurs when stakeholders at different levels of the educational system hold compatible and synergistic visions of learning goals and the means to achieve them. Imagine, for example, if an external assessment asked students to engage in problem-solving or writing tasks that were extensions of work they had done in class or if teachers received feedback from such examinations linked in a substantive way to their local curriculum.

While arguing for conceptual coherence within and across levels, *Knowing What Students Know* also clarified that a shared model of learning would not mean identical assessment specifications across different levels of the system. Rather, models of learning could be much broader at higher levels of aggregation to monitor trends or for program evaluation but would need to be more fine-grained to guide teaching and learning at the level of classrooms.

Big and Little Theories

Using theory to develop instructional and assessment systems requires an understanding of both general theories regarding the nature of human learning and development and of particularized, discipline-specific models of learning in practice (see, e.g., National Research Council, 2000). *Knowing What Students Know* identified four big or "grand theories," which they called perspectives: differential, behaviorist, cognitive, and situative (or sociocultural). Each of these grand theories makes quite different assumptions about the nature of knowledge and transfer; about the means for coming to know; and about the nature and role of interest, engagement, and motivation in learning. As a consequence, these theories also offer quite different conceptualizations of effective instructional interventions and related assessment tools.

The differential perspective focused on individual differences in mental abilities that were assumed to be fixed and predictive of future learning, rather than being an outcome of prior learning. From this perspective, measured differences were then to be used to assign individuals to instructional tracks corresponding to their different ability levels. The behaviorist perspective holds that learning occurs through the accumulation of stimulus-response associations, which can be reinforced by the administration of external rewards and punishments. Much has been written about the influence of behaviorism on the history of testing, the main concerns being its tendency to decontextualize and decompose content learning into tiny bits, its reliance on extrinsic motivation, and its inattention to thinking and reasoning processes. Cognitive theories, in fact, provided an answer to behaviorism by focusing on learning as a process of sense-making, that is, how individuals develop knowledge structures, construct mental representations, and in turn access these resources to answer questions, solve problems, and develop new understandings. A limitation of cognitive theories is that they focus only on what goes on inside the head of the individual learner. Like behaviorist theories, cognitive theories treat social interactions as contexts for learning, and not as helping to constitute what individuals are able to know, do, and become. They also ignore the wider historical and cultural contexts in which these interactions take place and that shape what kinds of knowledge are judged worthy of knowing or testing.

In the next section, we take up sociocognitive and sociocultural theories as the more compelling general theories to be used to design coherent curriculum, instruction, assessment, and teacher professional development. This is not to say that all aspects of prior theories are to be discarded. Reinforcement theory from behaviorism, for example, still works for some behavioral change efforts, and the concept of cognitive structures can still be used to think about how individuals make sense of new information. But these theories are insufficient by themselves to explain how human beings become more adept at thinking and doing. For this reason, more up-to-date and authoritative summaries of research in the learning sciences recognize the extent to which cultural practices and intrapersonal and interpersonal dimensions of learning are completely entwined with cognitive development (National Research Council, 2012b). Thus we argue that social models of learning are to be preferred because they provide a more complete account of how meanings, purpose, values, and motivation are jointly developed as part of deeper learning.

"Local" or small theories sit within these larger theories that inform their development. When *Knowing What Students Know* called for models of cognition and learning as the first corner of the assessment triangle, they had in mind more particularized models of how learning occurs rather than grand theories. Behaviorism requires the specification of detailed learning hierarchies, for example, because of its assumption that mastery of more advanced knowledge proceeds by the accretion of elemental, prerequisite learnings. Cognitive scientists were interested in mental processes, but early disappointments with teaching generic reasoning skills in knowledge-free domains led the field to focus more productively on cognitive processes involved in developing subject-matter–specific knowledge structures (Glaser, 1984). Wearne and Hiebert (1988), for example, studied "local theories" examining how children become proficient with the use of decimal fraction symbols. Situative perspectives have both drawn attention to and developed evidence related to "local instructional theories" for supporting learning trajectories in the context of local classroom communities (Cobb, 2000; Gravemeijer, 1994). They emphasize that such trajectories are made—and not discovered—by designing classroom contexts that are organized around disciplinary norms and practices that are integral components of the learning trajectory (Lehrer & Schauble, 2015). The *teloi* or aims of these learning trajectories are similarly not natural givens but articulations of who and what students should become.

Sociocultural Learning Theory: The Significance of Participation and Identity for Equity

All learning is fundamentally social, involving the individual's use of shared language, tools, norms and practices in interaction with his or her social context. It's not that behaviorism doesn't work by rewarding or "incentivizing" desired behaviors, but stimulus-response associations have not been found to lead to deeper understandings, complex reasoning, or knowledge use beyond initial training contexts. The sense-making focus of cognitive models is still a valued aspect of contemporary learning theory; but cognitive theory is limited to the extent that it considers only what goes on in the mind of individuals and requires separate theories of motivation to explain effort and investment in learning. Cognitivists have previously recognized social influences on learning, but sociocultural theory goes further in acknowledging how it is that one's cognitive development and social identity are jointly constituted through participation in multiple social worlds of family, community, and school. For the sake of brevity and clarity, we focus here primarily on sociocultural theory, acknowledging that there are actually several variants of these theories that attend to the social dimensions of learning: situative, social-constructivist, cultural-historical activity theory, and sociocognitive. As we explain later, sociocognitive models of learning are more limited than sociocultural theory, but they provide the more appropriate grand theory for understanding learning progressions.

Following many other contemporary interpreters of Vygotsky's work, we view learning as *the transformation of participation in valued sociocultural activities that are themselves changing* (Holland & Lave, 2009; Lave, 1993b; Lave & Wenger, 1991; Rogoff, Baker-Sennett, Lacasa, & Goldsmith, 1995; Rogoff et al., 2007; Rogoff, Paradise, Arauz, Correa-Chavez, & Angelillo, 2003). Participation in sociocultural activity necessarily involves more than simply acquiring knowledge; it involves processes of identification that, in turn, present opportunities for participants to become certain kinds of people in activity (Lave, 1993a; Lave & Wenger, 1991). This more encompassing view of learning became possible when anthropologists and sociologists joined cognitivists and developmental psychologists and importantly began to study learning in contexts outside of school. In real world settings, so-called cognitive learning and identity development are inextricably connected. Street vendors (Saxe, 1988) and basketball players (Nasir, 2000), for example, develop repertoires of practice and a sense of who they are as members of their community that are inextricably tied up with their knowledge of mathematics called upon in each setting. In this way, engaging in purposeful activity leads to one's increasingly proficient contributions to a community of practice without having to be separately or artificially induced.

Sociocultural theory offers a powerful, integrative account of how motivational aspects of learning—such as self-regulation, self-efficacy, sense of belonging, and identity—are completely entwined with cognitive development. Unlike the social psychological and developmental psychology literatures, where each of these variables and related interventions are studied separately, sociocultural theory helps us see how knowledge and self-beliefs are jointly developed in communities of practice, and correspondingly how sense of self and meaningful participation may be harmed in unsupportive learning environments where children may be positioned as unable or deficient learners.

Sociocultural approaches make it possible to design for equity in educational settings by attending both to who learners are when they join a community and who they might become. That is, they consider what is at stake for learners when they invest in developing new knowledge or skill in practice in a particular context. Countless studies have documented the lack of learning that occurs when students are treated as deficient or when members of non-dominant communities by race, language, or gender identity are asked to park their identities at the door to join the mainstream school or college culture. In contrast, sociocultural approaches such as funds of knowledge (Moll, Amanti, Neff, & Gonzalez, 1992) and cultural modeling (Lee, 1995) pay attention to students' everyday practices and find ways to connect "varied repertoires of practice" with academic disciplinary practices (Nasir, Rosebery, Warren, & Lee, 2014).

Going forward, providing young people with broad access to valued sociocultural activities is a key condition for new learning. A major challenge, however, is that Western societies purposefully segregate children from the mature activities of their communities by placing them in schools during work hours (Cole, 2010; Rogoff, 2003). School learning, as a consequence, becomes "encapsulated" or separated from other meaningful activities and subservient to larger societal goals of certifying the accomplishment of some (but not all) learners (Engeström, 1991). Rendering school learning meaningful and engaging students requires strategies for breaking down the barriers between school activities and mature sociocultural activities, for making visible, accessible, and personally relevant the knowledge, skills, and practices of those mature activities. It makes the learning that can go on in school more meaningful and purposeful.

Endorsing sociocultural theory as the most productive grand learning theory allows for the coherent co-design of curriculum, instruction, and assessment to enable deep learning over time. Curriculum frameworks (e.g., *A Framework for K-12 Science Education* [National Research Council, 2012a], which informed the development of the Next Generation Science Standards [NGSS Lead States, 2013]), now posit that proficiency involves much more than being able to recite core ideas of the discipline; it requires application of those ideas to explaining phenomena in the world and solving problems using the kinds of practices that disciplinary experts use. These efforts to break down the barriers between school activities and mature sociocultural activities are very different from earlier standards-based reforms because they are more thoroughly integrated and thus avoid the problem that sometimes occurred of doing "problem solving" only on Fridays. Later, when we take up specific assessment applications, we consider what means of giving feedback or certifying attainment for external audiences could be adopted that approximate mature disciplinary practices and allow learners to contribute to their communities in the process.

Discipline-Specific Models of Learning

In this section, we actually advance two arguments: first, that common curriculum materials coupled with teacher professional development are necessary to support equitable and ambitious teaching practices, and second, that such co-design of curriculum, instruction, and assessment requires discipline-specific models of learning (Pellegrino, 2014; Penuel & Shepard, 2016a). While it is true that many individual expert teachers create amazing, caring and challenging learning environments for their students, it is not possible working alone (even with open source materials on the Internet) to design coherent systems that build deep learning over time. Next-generation standards are discipline-specific and can serve as "curriculum frameworks" but they are too general to be curricula. A finer degree of granularity is needed to work out, not just what topics should be taught but also how practices and ways of thinking should be designed for as part of curriculum development. In a later section, we explain why we think that school districts and research-practice partnerships are the most likely institutional structures where horizontally coherent curriculum development and related opportunities for teacher learning could most likely be supported.

Fine-grained explication of learning goals and resources tied directly to frequent learning challenges are also essential for assessment design, especially to answer the important question of "what next?" Famously, Heritage, Kim, Vendlinski, and Herman (2009) found that "teachers are better at drawing reasonable inferences about student levels of understanding from assessment information than they are at deciding the next instructional steps." Formally designed curricula with embedded assessments based on a shared, detailed model of learning can serve as a resource for identifying frequent stumbling blocks as well as instructional moves that help to connect with students' current understandings. Foster and Poppers (2011), for example, described their work with teachers in the Silicon Valley Mathematics Initiative (SVMI) where they designed lessons to address learning needs identified by Balanced Assessment in Mathematics formative assessment tasks. Where teachers had previously responded to assessment results with reteaching or review sessions, the SVMI project sought to develop "reengagement lessons" for teachers to use in their classrooms. The idea behind lesson design was specifically to change how concepts were being presented based on insights from student work. Moreover, given the focus of professional development, new lesson strategies typically involved more student talk, so that, for example, even students with correct answers gained more practice in talking about part-to-part, part-to-whole, and ratio-to-fraction relationships. An additional benefit from using shared assessment materials was that anonymous examples of student work from across the project could be used to engage students in talking about why solution A or solution B might or might not be a good way to think about the problem.

Learning progressions are the most prevalent example of the more detailed models that Penuel and Shepard (2016a, 2016b) identified as sociocognitive models of learning. Other approaches include a "knowledge-in-pieces" or "facets" view of development (diSessa & Minstrell, 1998), which similarly identify ways to build on students' intuitive understandings as well as learning experiences needed to revise ideas that are productive in one context but problematic in others. Sociocognitive models attend to the social nature of learning and to discipline-specific ways that core ideas and practices are developed over time. The general "social" theory underlying sociocognitive development efforts is consistent with sociocultural theory in that it posits "that individual cognition develops through social interaction, as individuals solve problems, complete tasks, and devise strategies to pursue particular goals" (Penuel & Shepard, 2016b, p. 147). We resisted labeling the extensive literature on learning progressions as sociocultural, however, because existing research studies tend to build from common starting points rather than making students' interests and community experiences part of how interventions are designed. Moreover, they treat target ideas and practices as fixed end points, rather than changing. We consider later how sociocognitive learning models can be adapted subsequently in local communities of practice at the so-called implementation stage, even if local contexts were not part of the initial design.

Learning progressions are defined as "descriptions of successively more sophisticated ways of reasoning within a content domain" (Smith, Wiser, Anderson, & Krajcik, 2006, p. 1). They are different from either item-anchored test-score scales or curricular scope and sequence strands, because they require *both* conceptual analysis and empirical testing of hypothesized developmental pathways (Corcoran, Mosher, & Rogat, 2009). Although Rasch-scaled progress maps or progress variables were offered in *Knowing What Students Know* as early examples of learning progressions, taking learning theory seriously means that measurement specialists should not merely scale together samples of students who have and have not had effective instructional opportunities. Instead, learning progressions require an iterative research and development approach whereby the conjectured curriculum-specific learning theory—with related instructional activities and assessment tasks—is tried out and revised in response to evidence as to which curricular resources actually contribute to learning progress.

Six years after publication of their "possible or proposed" learning progression for matter and atomic-molecular theory (Smith et al., 2006), Wiser, Smith, and Doubler (2012) provided an extensive report illustrating the kind of programmatic research needed to develop and "validate" a learning progression's theoretical claims and its concomitant instructional and assessment tasks. Focusing only on the learning progression for matter (LPM) in the elementary grades, Wiser et al. (2012) first used existing research to inform LPM conceptualization and design of the Inquiry Project Curriculum, an elementary-grades curriculum that focuses on developing students' understanding of the nature of matter. For example, early childhood researchers know that many 2-year-olds know the word "heavy" (Hart & Risley, 1999) and that most children come to understand weight and comparative heaviness by hefting. Wiser et al.'s (2012) overall strategy was to find ways to "reduce the incommensurability gap between students' and scientists' knowledge networks progressively and coherently" (p. 380). Thus they designed activities, for example, to help third graders move from thinking about weight qualitatively to measuring it quantitatively. Their design process was quite complicated, however, as these concepts do not exist in isolation but, in fact, involve an interplay of weight, material, amount of material, size, and volume of material. Thus, the curriculum design required theorizing these intersections and then testing which combinations and revisiting of concepts best helped students revise their ideas and move toward more sophisticated understandings. Importantly, the curriculum project also followed up with classroom observations and post-interview data to evaluate the efficacy of specific instructional moves. For example, 60% of the treatment group who used a weight number line argued confidently that "even tiny pieces of clay must weigh something and must take up space" (Wiser et al., 2012, p. 388) as compared to 11% of controls. These data illustrate the type of evidence needed to support the instructional validity of assessments intended to improve learning (Pellegrino, DiBello, & Goldman, 2016).

The demanding process we are describing for developing and testing horizontally coherent curriculum, instruction, and assessment activity systems does not imply, once such systems are developed, that they should be rigidly implemented. Even when learning progressions or other curricular resources have a strong empirical basis, it does not follow that all learners in all contexts will progress reliably through a sequence as specified. This should serve as a cautionary tale for measurement specialists first venturing into the arena of curriculum and instructional design. To be used in ways more compatible with sociocultural theory, learning progressions should be thought of as cultural tools that will require further adaptation in local contexts (Lehrer & Schauble, 2015; Penuel, 2015). Assessment tasks or guiding instructional questions developed as part of a progression should be designed to elicit and illuminate student thinking so that the diversity of students' ideas is part of what is acted upon in instructional activity. That is, such tasks should not simply be designed to "recruit" students' interests into disciplinary ways of thinking, but rather to help them recognize those ways of thinking and relate them to their own (Bang & Medin, 2010).

Thus, sociocultural theory posits a role for assessment to elicit students' relevant interests and experiences and their identities so as to inform instructional practice. For example, at the beginning of a unit in the *Micros and Me* curriculum (Tzou & Bell, 2010), students take photos of things or activities that they do to prevent disease and stay healthy at home or in their communities. They then share these photos in class as a way to bring personally relevant experiences into the classrooms and draw connections to learning goals for the unit.

Purposes of Assessment and Levels of the System

Just as sociocultural theory and discipline-specific little-theories of learning are key contributions to assessment from the learning sciences, there are also foundational principles from measurement science that speak to the design of effective assessment systems. *Knowing What Students Know* (National Research Council, 2001) acknowledged the learning requirement for coherence across levels of the educational system, but at the same time attended to validity requirements, calling for differences in "tests" designed for different levels of the system. Citing Cronbach and Gleser's (1965) distinction between "fidelity" and "bandwidth" they explained the necessary "trade-offs" in assessment design between the fine-grained information needed by a teacher during a particular lesson or unit of study, as compared to the broad and more comprehensive summary information needed by policymakers once per year. These ideas about how assessments could be faithful to a shared model of learning but still be tailored to different purposes was also addressed by a subsequent National Research Council workshop report (2003). To serve their purposes well, large-scale assessments must ensure comparability, which typically means standardized administration procedures, uniform administration dates, and independent

Table 1. Assessment Purposes and Levels of the Education System With the Most Frequent Intersections Highlighted

	Assessment Purpose				
Educational System	Formative Assessment	Grading	Program Evaluation for Teachers, Schools, and District	Accountability	Large-Scale Trends and Research
Classroom	▨	▨			
District			▨		
State				▨	
National and International					░

Note: The four cells with dark highlighting are the four intersections of assessment purpose and level discussed in this article.

performance by students. In addition, for the sake of fairness, large-scale assessments can't favor one local curriculum over another, so they tend to use more generic representations of learning goals, and they have to be cost effective. In contrast, classroom formative assessment is needed at unique times, can sometimes involve assisted performance (as in dynamic assessment), and provides greater insight when tied directly to local curricula. When formative assessment is seamlessly integrated in instruction it involves no additional costs in dollars or instructional time.

Knowing What Students Know (National Research Council, 2001) called for alignment or vertical coherence "as one moves up and down the levels of the system, from the classroom through the school, district, and state" (p. 256). We use these levels of the education system and respective purposes to organize the remaining sections of the article. In Table 1, a simple cross-classification is shown of four levels of the education system crossed by five well-known purposes for assessments. Except for the first purpose, which is classroom formative assessment, the other purposes are summative in that they represent culminating, point-in-time pictures of student achievement. Note, however, that district assessments may also be used formatively to make curricular or programmatic improvements. The five shaded boxes in Table 1 show the purposes for assessment most often taken up by each of the levels, respectively; the first four of these are the specific applications we propose to consider next. Teachers (and students) in classrooms are most responsible for formative assessment to support learning. Classroom assessment must also support grading purposes, which could sometimes be at cross-purposes with formative assessment. This incoherence and tension could be mediated if more explicit thought could be given to the relationship between ungraded formative observations of performance and formal grading and reporting requirements.

Districts most often use district-level assessment results to track trends, compare schools, and evaluate programs. But districts could also sponsor assessment materials to be used by teachers and schools; and they could more explicitly attend to the unintended but demotivating effects of constant grading and reporting requirements experienced by teachers and students. States are responsible for state tests used for accountability. Clearly these are not all of the possible purposes for assessment in school systems —special education placements and teacher evaluation are two obvious omissions—and we do not claim that there are no applications in the off-diagonal cells. For example, districts often use state test results for program evaluation, as well as for state-sponsored school accountability. The last row and column are included for the sake of completeness but are not addressed in this article. National and international assessments such as the National Assessment of Educational Progress (NAEP) and the Programme for International Student Assessment (PISA) are used to monitor large-scale trends and for research. Of course, state and district testing programs are also used to report on trends and for research. We address these four particular applications because they are the most prevalent and because they are the most critical if assessment is to be used in coherent ways to improve student learning.

Given our argument previously that shared curricula coupled with teacher professional development are necessary to support equitable and ambitious teaching practices, our analysis of assessment applications must also consider a third dimension having to do with *curriculum specificity*. In their studies addressing instructional sensitivity, Ruiz-Primo and colleagues (Ruiz-Primo et al., 2012; Ruiz-Primo, Shavelson, Hamilton, & Klein, 2002) identified five different assessment "levels" or "distances" based on their *proximity to the enacted curriculum*. This idea about distance from the classroom is correlated with level of the education system but is more about *similarity* or the *closeness of the match* between assessments and specific instructional experiences.

As suggested above, curriculum specificity is necessarily abandoned in large-scale test design when a district or state wants to compare the performance of students across jurisdictions using quite different curricula. But such tests then are not of much use for detailed formative feedback. There is also the issue that as assessments become more decontextualized relative to the original curriculum and instruction, student performance may vary in ways that would be predicted by sociocognitive theory. There is likely to be a shift across levels in the degree of social context and mediation of the assessment activities and in the degree of remoteness or abstraction from the original processes of learning and instruction with respect to the tools, artifacts, discourse structures, and persons involved. As noted by Ruiz-Primo et al. (Ruiz-Primo et al., 2002), then, curriculum-general tests could underestimate the learning gains documented by more instructionally relevant assessments.

To design for coherence across levels of the system, we address this issue of curriculum specificity in the next, applications sections of the article, paying particular attention to the issue of learning transfer. *Transfer of learning* refers to the use of knowledge learned in one context in some other context. Following Hickey and Pellegrino (2005), we acknowledge that the conception of transfer depends on assumptions about knowing and learning, which differ across learning theories. Interestingly, however, everyone believes in transfer, that is, they want what is learned in school to be useable in the outside world. But differences arise depending on one's beliefs about what instructional methods will ensure transfer and what evidence should be accepted as proof of transferable learning.

From a sociocultural perspective, the ability to use one's school knowledge and ways of participating are directly connected to the real world to the extent that instructional activities have already been designed to engage students in age-appropriate practices, such as modeling and argumentation, leading to full participation in mature practice. Learning progressions provide a tool for thinking about and evaluating whether such visions of learning progress actually work as intended. Notice that disciplinary practices included in present-day standards represent modes of inquiry and ways of developing understandings that are expected ultimately to generalize or "transfer" beyond the boundaries of specific content units of study, such as functions in mathematics or heredity in science. However, practices may or may not transfer across content domains because it is not just about modeling or explanation or argument but also about the substance of what is being modeled. In elementary school, children who have been taught to draw pictures to represent fractions will regularly do so—hence we might claim transfer of this practice, yet novice learners might have difficulty if their local curriculum used fraction strips instead of pie charts. Similarly, at the high school level being able to draw or interpret a phylogenetic tree and being able to use such a model to demonstrate one's knowledge of evolution will very much depend on whether this specific representational format as well as necessary content had received attention in local curricula.

Thus, developing curriculum and related assessments requires that subject-matter experts and measurement specialists explicitly consider how much "sameness" is required for each assessment purpose. Surely new learning requires multiple opportunities with familiar tasks and routines, but teaching for "robust" understandings (Shepard, 1997) also requires extensions and exposure to other ways of thinking about a problem (National Research Council, 2012b). In the next sections we consider how formative assessments should contribute to learning that is later graded (when grading is required) and how districts could design curriculum-based assessments to support learning in classrooms. When external assessments can't be co-designed with curriculum-based assessments, we consider how these other genres could be explicitly addressed when they land in classrooms.

Formative Assessment as Part of Ongoing Instruction in Classrooms

Bennett (2011) offered a succinct summary regarding the controversy surrounding the definition of formative assessment. One camp thinks that "'formative assessment' refers to an instrument (e.g., Pearson, 2005), as in a diagnostic test, an 'interim' assessment, or an item bank" (p. 6). The other camp, quoting from Popham (2008), holds that "formative assessment is not a test but a process" (p. 6). Bennett went on to say that this dichotomy reflects an oversimplification; but, given this rhetorical surround, it is useful to clarify our position. This article is written from the learning research perspective of the process camp. This does not mean, however, that we would rule out ever using closed-form item formats nor do we rule out quantifications, especially those representing gains on substantively grounded learning progressions. Later, especially when we talk about district level uses of data, we acknowledge that quantitative summaries are important and useful; moreover, artifacts that students generate in this context could directly feed back to support formative assessment as a process. Measurement experts can play a particularly important role in building vertically coherent systems by finding ways to establish linkages between rich instructional tasks and aggregate achievement indices. At the same time, measurement colleagues venturing into classroom learning contexts should be mindful of the harm and distortion that occur when assumptions about standardization and quantification are brought into classroom processes unnecessarily.

For the purposes of classroom teaching and learning we contend that test-format data systems have been given too large a presence in today's classrooms. This is problematic because both their representation of learning goals and the information they provide are limited, and their use often has negative effects on student identity development and motivation. Our brief argument here, in favor of curriculum-embedded-in-the-process-of-instruction formative assessment, makes two main points, focused first on the usefulness of the "information" provided by assessment activities, and second on the ways that materials and tools support productive, learning-focused "assessment cultural practices."

Instead of "scores," which offer teachers little information about what to do next, it is much more important that formative assessment questions, tasks, and activities provide instructional "insights" about student thinking and about what productive next steps might be taken. Quantitative data systems frequently provide teachers with results in the form of a grid showing class rosters crossed with item or subtest scores. Teachers then typically use these data to reteach the standards or objectives missed by the most students (Foster & Poppers, 2011; Shepard, Davidson, & Bowman, 2011), and they identify students who missed the greatest number of items for special tutoring or after-school help. In these scenarios, reteaching efforts are not adjusted based on information from the assessment, although sometimes teachers say they make changes just to try something different. Given what we know about how such interim assessments are constructed, it is not surprising that identifying standards not-yet-mastered does not give teachers access to student thinking. To intervene with an individual student, a teacher still needs to follow up with more individualized student conversations, which is generally not possible due to time constraints. We do not claim that reteaching never helps but rather, because of the vagueness of the treatment, that its benefits are likely to be small in comparison to the instructional time lost. In addition, typical interim tests offer students an impoverished vision of intended learning goals.

Specific curricular tasks used as part of instructional conversations can provide more actionable information and also provide insights about what a student does know that can be used as a basis for addressing understandings that are still out of reach. In the fraction problem illustrated in Figure 1, for example, this student might be asked to talk aloud about how they solved one of the correct pathways; then, in a safe and trusting space, they might be supported to talk through and rethink a solution for one of the "hard number" pathways. Note that sharing explanations can be done with the whole class and does not require one-on-one time. Even sharing "incorrect" solutions can be helpful, so long as a classroom culture has been established that is mutually supportive. In the data set from which this example was drawn, many

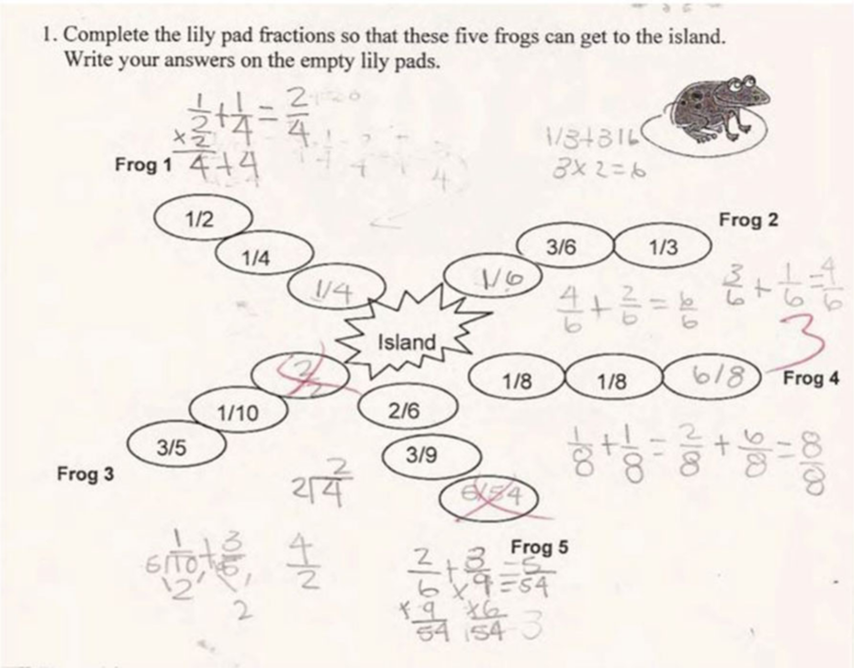
1. Complete the lily pad fractions so that these five frogs can get to the island.
Write your answers on the empty lily pads.

FIGURE 1. An example of student work illustrating understandings to be built upon plus specific concepts and skills still in need of support.[1] [Color figure can be viewed at wileyonlinelibrary.com]

students also drew pie charts or bar graphs to think about a set of Leapfrog Fraction problems; these would also be valuable resources to share in class discussions.

The informational value of assessment items or tasks also matters in terms of the feedback provided to students. It does little good to tell a fourth grader that his essay is a "2" or that he needs three more items correct to reach mastery on a numeracy test. Feedback is more likely to be beneficial when it helps students see *how* to improve (Black & Wiliam, 1998), and this invariably calls for information that is qualitative rather than quantitative. Talking to students about their scores—and implicitly representing learning goals numerically instead of substantively—has motivational consequences as well. It leads to the commodification of learning and promotes extrinsic over intrinsic motivation, which is antithetical to what Sadler (1989) had in mind when he argued that formative assessment should foster development of expertise by helping students to internalize the features of good work.

The empirical literatures on both feedback and intrinsic motivation are vast, but they lead to quite similar conclusions. Feedback focused on features of the task or a student's strategies most often leads to positive learning gains, while feedback focused on the person in comparison to others actually harms learning (i.e., produces negative effect sizes) (Hattie & Timperley, 2007; Kluger & DeNisi, 1996; Shute, 2008). Feedback conveyed in the spirit of improving also helps to foster a "learning orientation" whereby students work to feel an increasing sense of mastery and of becoming competent. Conversely (or perversely), a classroom focus on extrinsic motivations develops in students a "performance orientation," aimed at getting good grades, pleasing the teacher, and appearing competent. These motivational constructs dating from the research in the 1980s are still valid, though today involve much more complex theorizing by social psychologists regarding identity development and school engagement (Eccles 2009; Wang & Eccles, 2013). We can also understand these patterns in terms of sociocultural theory, that is, where students may be positioned adversely or, as valued participants, may willingly devote effort to become increasingly able to contribute.

When we talk about desirable and productive "assessment cultural practices," we are referring to the social norms and meanings associated with assessment processes in classrooms. A significant contribution of work by Black, Harrison, Lee, Marshall, and Wiliam (2003), and the formative assessment literature more generally, has been to change the social meaning of assessment and to refocus it on *learning*. In learning-focused classrooms, students are willing to take risks to learn from a trusted adult and classmates. As with feedback, there are extensive literatures on *self-regulation, self-assessment,* and *peer-assessment.* These formative assessment processes contribute both to metacognition and identity development because they involve students in activities to internalize the criteria for judging quality work, take responsibility for evaluating and seeking to improve their own work, and in so doing take on the role of disciplinary expert. As noted by Shepard, Hammerness, Darling-Hammond, and Rust (2005), "Internalizing what criteria mean in a particular discipline is not just about learning the rules for grading—it literally means learning the discipline itself" (p. 298). From a sociocultural perspective, cognitive and motivational dimensions are completely blended as students are supported in becoming a certain kind of person—for example, a writer, a curious investigator of nature, or a data-modeling expert—in the context of engaging in disciplinary practices.

What we are trying to do in schools is to construct learning environments that have the same sense of common purpose as in real world settings. In more authentic contexts, assessment and critique are a natural part of joint work. Penuel and Shepard (2016a) used the example of Youth Radio, where youth have real responsibilities for producing programming for external audiences. At each phase of production, from initial decisions about what makes a "good story" as well as technical decisions regarding music, voiceover, interview footage, and so on, youth engage in informal assessments where they become increasingly adept at using the criteria of the genre. Postproduction, formal critique sessions are held, and editing (the ultimate evaluative process) conducted in full public view, is accepted as inherent to the work itself. The practices of evaluation of performances and products, moreover, resemble evaluation practices that occur within professional practice, in this case of radio production (e.g., Soep, 2006).

In the next section, we take up the topic of grading, which more than any other single factor tends to undermine the good intentions of formative assessment. Although we say more later, a key principle to be emphasized is that formative assessments should not be scored or graded.

Grading

Brookhart et al. (2016) provide a comprehensive review covering 100 years of research on grading. Countless surveys of teachers' grading practices document the extent to which teachers reward what McMillan (2001, p. 25) calls "enabling factors" (effort, ability, improvement, work habits, attention, and participation) in addition to mastery of learning goals. These factors and the extent to which they vary among teachers, along with the vagueness of criteria used to evaluate achievement outcomes, largely explain the notorious unreliability of teachers' grades. Ironically, however, the composite nature of grades, averaged across many teachers, as well as their "multidimensional" character as described by Brookhart et al. (2016) are what make grades a better predictor of future success in school than standardized test scores (Bowers, 2010; Brookhart et al., 2016). This measurement literature, focused on grades as an indicator of achievement, does not address the effect of grading practices on learning. As we examine here, evidence about learning consequences comes instead from the motivational literature cited previously.

To be coherent with formative assessment, a critical requirement is that grades be based on the same learning goals toward which instructional activities and formative feedback are aimed. This may seem obvious, but in fact it is still the case that many classroom quizzes and unit tests rely on formats that make them an impoverished rendition of learning activities. Curricular resources can be a profound help in this regard especially if they explicitly consider the relationship between instructional activities, formative questions, and culminating or transfer tasks. Good examples are the National Writing Project's College-Ready Writers Program mini-units and formative assessments designed to help teachers help students write more compelling source-based arguments (Gallagher, Arshan, & Woodworth, 2017). Students received specific practice and feedback regarding their abilities to develop a claim; connect evidence to a claim; select relevant evidence from source material; comment on the credibility of source material; and use source material for purposes of illustrating, invoking authority, as a take-off point for extension, or as a source of disagreement

In the same vein, grading practices are the most consistent with formative assessment cultural practices—and make the most sense to students—when the focus is consistently on student learning. A learning or mastery focus argues against "extra-credit" points, for example, intended to compensate for earlier learning that fell short, but argues instead for integrative culminating assignments that could replace an earlier assessment once students are able to demonstrate more complete mastery. Point systems in general and electronic systems designed to keep parents continuously informed tend to work against the logic of formative assessment because they assign points as if early steps in learning are "finished" rather than providing a substantive and still changing picture of developing competence. Research on effective feedback argues against grading because the grade itself becomes the focus of attention rather than attending to the means for improvement (e.g., Butler & Nisan, 1986).

Many teachers are uncomfortable with these findings, believing that they must use points to "motivate" students and control behavior. (Others grade because they are held to account for doing so by school and district policies.) Of course, we should recognize that rewards do work at some level. For example, more students will bring a pencil and their textbook to class if they know it will affect their grade. It also seems reasonable to "incentivize" things like effort and participation, because they are known to affect learning. Yet, the motivation literature is quite clear that making these prerequisite behaviors dependent on external rewards harms learning, especially when the rewards are seen as controlling. The conclusion that "extrinsic rewards drive out intrinsic motivation" is based on a large body of controlled experiments (Deci, Koestner, & Ryan, 1999) showing that subjects of all ages, preschool to college, show less interest and less engagement in target activities, after being rewarded for them and then having rewards withdrawn, than they had demonstrated prior to the beginning of such interventions. Instead of using grades as an incentive system, the alternative is to develop a classroom culture where students are able to experience the internal "rewards" that come from meaningful participation, a sense of increasing competence, and exploration of new ways of being, knowing, and acting in the world.

The argument is also made that point systems and regular grading of formative assignments are necessary to keep parents informed. Plus, there is the issue that students gain deeper insights about the meaning of disciplinary criteria and scoring rubrics if they have the opportunity to engage with them in the context of their own work and with shared group projects. Still, there is ample evidence that putting a grade on a paper detracts from the value of formative feedback, and there is not one best way to resolve these tensions. Self- and peer-assessment offer one means to help students gain experience with the meaning of criteria without formally recording grades. These processes also contribute to a positive assessment culture, so long as class discussions stay focused on substance and do not devolve into haggling over points. As noted previously, point systems—and the familiar grading practice of averaging together points from preliminary, intermediate, and culminating assignments—work against the idea of learning from early efforts. As an alternative, Shepard et al. (2005) suggested using "as if" grades (p. 305) to help students see the meaning of criteria but also to signal that such judgments are temporary. In general, to be consistent with a productive formative assessment culture, instructional activities and required assignments should afford students ample opportunity to use insights from formative assessment to improve the quality of their work. It is these culminating performances that should be reflected in summative grades.

Standards-based grading[2] systems are increasingly promoted by professional development leaders (Guskey, 2009;

Marzano, 2010; O'Connor, 2002) as a way to refocus grading on achievement and to have grades more accurately reflect mastery of intended learning goals. These important ideas about reporting learning in relation to well-articulated, substantive performance targets date from prior standards-based reform efforts in the 1990s (Wiggins, 1996) and even from earlier ideas about criterion-reference testing (Glaser, 1963). Clear standards and criteria support learning by drawing explicit links between formative and summative assessments. In addition, comparing a student's work to performance goals instead of normatively to classmates avoids many of the negative effects of grading found in the motivation literature. Measurement specialists should be aware of these underlying principles so as to forestall the distortions that have occurred in the past as, for example, when the idea of criterion-referencing was shifted from clearly articulated criterion performance to cut scores or 80% definitions of mastery. Likewise education leaders might need to be cautious about whether mandated standards-based systems are likely to be effective. Brookhart (2011) wisely makes the case for the kinds of professional conversations that are needed to think through the audiences and purposes for grading and how these should, in turn, guide grading reforms. Our advice would be to consider grading practices in the larger context of standards implementation, formative assessment, and subject-area-specific professional development.

District-Level, Curriculum-Based Assessment

Districts have their own needs for assessment data, for purposes such as school accountability and program evaluation. In this section, however, we focus on our main argument, namely that districts could be the most appropriate level to design and implement *curricular activity systems* (Roschelle, Knudsen, & Hegedus, 2010) to address *both* district-level and classroom-level assessment needs. Classroom assessment needs encompass both curriculum materials and the system of supports for teachers and students designed to promote effective implementation of those materials. To support student learning, curricular activity systems are grounded in contemporary theories of learning and are designed to coherently integrate curriculum, instruction, and assessment as well as related teacher learning opportunities. Although technically designing for coherence is possible with behaviorist or cognitive theories, we have argued here for a sociocultural theoretical perspective because it best takes account of the social nature of cognitive development and, in addition, attends to affirming, identity-producing learning opportunities that come through participation in communities of practice. It also requires us to attend to the ways that these curricular activity systems themselves change over time, both as students, teachers, and educational leaders respond to challenges that arise from implementation, and as they come into contact with larger organizational and institutional pressures. As such, they remind us that coherence is not a static "property" of the relations among curriculum, instruction, and assessment but something that local actors must continually "craft" (Honig & Hatch, 2004).

Our thinking in arguing for districts as the locus for curriculum development (Shepard, Penuel, & Davidson, 2017) comes from the key role played by districts in implementing standards and providing teacher professional development. Districts also play a key role in promoting and monitoring equity with respect to opportunities to learn. Developing coherence and ensuring equity requires that districts develop a vision for high-quality instruction and then build the commitment and capacity of teachers to enact and evolve that vision (Cobb & Jackson, 2012; David & Shields, 2001; Supovitz, 2006). We also note that the district level is where decisions are made to allocate resources for textbooks and interim tests. In local control contexts especially, a district, a group of districts, or a research-practice partnership might be better positioned than states to attend to the coherent integration of curriculum, instruction, assessment, and teacher professional development.

We argue, then, that sociocultural theory broadly, and discipline-specific models of learning, could be the basis for developing a horizontally and vertically coherent curriculum with embedded assessments of three types: formative assessment activities to surface student thinking and further learning within instructional units, summative unit assessments used for grading that explicitly address transfer and extensions from previous instructional activities, and district-level assessments designed in parallel to unit summative measures but with particular attention to program-level evaluation. These three levels of assessments should be substantively congruent, but they require different levels of quantification, technical rigor, and comparability across units.

In the Inquiry Hub research-practice partnership between Denver Public Schools and the University of Colorado Boulder, we are working toward creating a coherent and equitable curricular activity system in secondary science classrooms modeled on these principles. The partnership operates as a design research partnership (Coburn, Penuel, & Geil, 2013) whose mission is to design, test, and implement tools and strategies for supporting teachers in developing ambitious and responsive instruction to help all students achieve at high levels in mathematics and science. For the past 3 years, the partnership has been focused on developing and testing curriculum materials for high school biology that are aligned to the Next Generation Science Standards and that are anchored in explaining phenomena and solving problems and organized around coherent "storylines" (Reiser, 2014). Designing formative assessments, unit assessments, and district assessments that reflect the vision of science proficiency embodied in the standards has been an integral part of this effort.

Our formative assessments elicit students' understanding of disciplinary core ideas and crosscutting concepts in the context of science, and they seek to make visible the ways students' own interests and experiences relate to phenomena being studied. For example, the opening of each unit begins not with a formal "pretest" but rather with the presentation of a phenomenon—a video of boys living with Duchenne's Muscular Dystrophy, for example, in a unit on genetics—and asks students what they notice and what they wonder about. Students draw initial models to explain what they see, and they generate further questions that arise from discussing gaps in their models. The teacher invites them to bring up related phenomena in their everyday lives—relatives with inherited diseases—and add questions for the class as a whole to investigate. Then, the class comes together to cluster, prioritize, and define ways to investigate their questions. This sequence of activities helps put students' questions and ideas at the center of instruction, and because it is repeated across multiple units, students begin to get a "grasp of practice" (Ford, 2008) in asking questions in science, within the context of specific disciplinary core ideas. Moreover, it inducts students into the practice of developing and using models to explain phenomena, creating both a need and desire to learn disciplinary core ideas.

Additional formative assessments embedded in units engage students in collaborative model evaluation and revision that provide teachers with evidence regarding students' experience of classroom activities. For example, as students pursue answers to the questions they generate, they build models incrementally—each lesson adding a piece to their model of the phenomenon at hand. In lessons focused on putting model pieces together, students go public with what they've learned, sharing models using a gallery walk format with classmates, reviewing what is common and what is distinctive in different groups' models. They offer critiques to one another on the basis of what evidence their models account for, as well as what is missing, and then student groups revise their models on the basis of peer feedback. Periodically, teachers also use an electronic exit ticket that includes opportunities for students to say what they figured out related to the phenomenon that day and to evaluate whether the lesson as enacted helped them answer at least one of the questions the students have decided to pursue (Penuel, Van Horne, Severance, Quigley, & Sumner, 2016). These exit tickets also ask students to report on how they learned that day, whether they contributed ideas to the group, and their emotional experiences. Teachers can use these data to compare facets of core ideas that students were supposed to learn from the lesson in the storyline, and they can also examine patterns of participation to diagnose inequities of experience and variability in their own success in supporting students' progress toward explaining phenomena.

Each unit is comprised of two to three anchoring phenomena, where each of the formative assessment types described above supports the building of a more generalizable understanding of a disciplinary core idea and a more sophisticated grasp of the practice of developing and using models to explain phenomena. Unit assessments not only ask students to build explanatory models of each phenomenon that describe key components, interactions, and mechanisms but also to identify similarities between phenomena with respect to key mechanisms. In our genetics unit, for example, students build a model of how genetic diseases are inherited and inhibit (in some cases) the production of proteins that matter for key functions in the body. Then they apply these understandings to build an even more sophisticated understanding of genetics through the study of powerful emerging gene editing technology, CRISPR-Cas9 (Doudna & Charpentier, 2014). In this second part of the unit, students tackle new questions about the ethics of genetic engineering and gain practice with a new form of dialogue—a World Café—organized to support civil, evidence-based, democratic discussion of socio-scientific issues. In class, the summative assessment requires students to describe how the CRISPR-Cas9 system might be used to cure genetic diseases and the mechanism by which it could work. They also prepare for a performance in the community of the World Café structure where they act as co-facilitators (with adult leaders) of peers, parents, and teachers in discussions related to the ethics of genetic engineering. This provides an opportunity for young people to connect their classroom learning explicitly to the community and apply their understandings of both how to structure productive dialogue and of genetics to contemporary socio-scientific issues.

We are involved in developing district-level assessments that cohere with these curricular goals as well. To some degree, this is born out of necessity: teachers tell us all the time that unless the district assessment focuses on these goals, they have too little "cover" from administrators in their building to try out new curriculum materials. After all, Colorado has not yet adopted the Next Generation Science Standards, and there are competing (though not necessarily misaligned) frameworks that teachers' principals use in observing teachers' classrooms to evaluate their performance.

The district-level assessment re-design encompasses both student accountability and teacher evaluation. We have worked with teachers both inside and outside our curriculum design efforts to develop extended, multicomponent tasks that assess "three-dimensional" science proficiency, reflecting integrated understandings of disciplinary core ideas, science and engineering practices, and crosscutting concepts. These tasks are incorporated into district-administered unit assessments, whose timing is aligned to the pacing guides provided with our co-designed curriculum units. We are in the process, too, of developing validity and reliability evidence for these extended tasks, so that we can eventually use them to help us evaluate the effectiveness of the curriculum materials. In addition, we have developed qualitative, multilevel rubrics for assessing the quality of students' models of phenomena that teachers can use to develop a portfolio of evidence of their own effectiveness, according to the district's Student Learning Objectives (SLO) process. These rubrics are tied to hypothetical learning progressions keyed directly to Appendix F of the Next Generation Science Standards, which focuses on how students' grasp of the practice of developing and using models is expected to progress. Similarly, our own lesson-level objectives are keyed to this Appendix, which helps to promote coherence.

Because our curriculum is still in development, and because the external environment to our partnership (which includes parts of the district) is constantly shifting, the coherence among these different assessment levels is something we must constantly recraft. Unit launches are revised to reflect better what students' questions are when phenomena are presented, and when revising units entirely new phenomena are sometimes selected to better reflect students' interests. Model building and revision activities are changed to better support equitable participation in classroom discussions, and multicomponent task prompts are revised to elicit students' understandings. Perhaps most importantly, professional development activities are created and revised to support teachers in making shifts toward equitable, three-dimensional science teaching and learning.

State-Level Assessment, Standards-Based Assessment, and Accountability

The primary focus of state-level assessment is district- and school-level accountability. If states had curricular responsibility, then what we have described as possibilities for district-level, curriculum-focused assessment designs would also be possible at the state level. More typically, however, in states with local control, state-level assessments are based on standards frameworks, which are only the shell for what might become curricula. Without a shared curriculum, states most often relinquish the idea of coherence with classroom- and district-level assessments by building more generic tests.

In the past, it might have been possible to develop such "curriculum-free" state tests that were an adequate fit with local curricula. As next-generation standards become more ambitious, however—calling for complex demonstrations of proficiency and the integration of disciplinary practices and content—it becomes increasingly difficult to be fair across curricula and still detect learning gains from those respective curricula. In the past, a close look at international comparison data taught us that country rankings could change dramatically by subtopic, item, and item type (Schmidt, Jakwerth, & McKnight, 1998), supporting the point that match to curriculum matters. Today, alignment or coherence is considerably more challenging because local curricula cannot possibly attempt to cover all possible intersections of disciplinary content and practices at a given grade level. So who will make these choices to ensure deeper instructional opportunities, and how will they be aligned across levels of the system?

It is beyond the scope of this article to try to generate possible solutions to this dilemma, but we should recognize some plausible pathways. One path, taken by PARCC and Smarter Balanced, is still to build curriculum-general tests, which do not call upon the particular tools and activities with which students engage in specific curricula. Another is illustrated by the experiment in New Hampshire, which places greater authority in district level assessments but attends to the issue of comparability across schools and districts (roughly) by interspersing administrations of common performance assessments and Smarter Balanced at key grade levels along with local performance assessments. Note that these assessments do not claim to be equated. Rather the multiple-assessment design is intended to check on the comparability of claims made for policy purposes about percent of students who are proficient. The New Hampshire example assumes, as we have here, that the district is the appropriate level for co-development of curriculum, instruction, assessment, and teacher professional development.

In cases where districts lack the resources to design curricula to help teachers engage with next-generation standards and ambitious teaching practices, states could serve as a resource by providing "replacement units" that model the desired instructional activities with embedded formative and summative assessments (Marion & Shepard, 2010). Replacement units are also a strategy that states might use, short of designing entire curricula, as a means to support standards implementation without top-down mandates. In the National Research Council report (2014) on developing assessments for Next Generation Science Standards, several options are discussed—such as performance assessments, portfolios, and projects—whereby state level monitoring assessments could tap into more complex forms of knowing consistent with a sociocognitive framing and the intentions of new standards. Such next-generation assessment designs would enable greater coherence between state and local reform efforts rather than perpetuating unwanted coherence such that what gets tested limits what gets taught and how it is represented in classrooms.

Conclusions

The kinds of equitable and coherent systems of assessment we are calling for that are organized around a common, sociocultural vision for disciplinary learning and where curriculum, instruction, and assessment are closely coordinated represent an extraordinarily ambitious goal for education. Such systems will require intensive, ongoing work to build and maintain. And they will—because of the political nature of our educational systems—always be precarious and subject to challenge. Nonetheless, local efforts to build them within subject-specific subparts of large systems like the Inquiry Hub partnership give us hope that such systems could be developed.

These remain isolated cases, in our view, because it is difficult for districts to allocate the time and resources at their own disposal for deep and iterative cycles of development of curricula and embedded assessments, especially in all subject areas at once. We see a strong need for networks of districts and states to come together for this purpose, pooling resources, to engage different stakeholder groups in their communities in this joint work. This would require in many cases re-allocating resources devoted for textbooks and instructional technology purchases to invest in educators' time to build Open Educational Resources and participate in endeavors where they have opportunities to learn how to design materials and assessments, side by side with professionals who specialize in curriculum and assessment development. Engaging these outside partners will require funders to invest in partnerships, too, both to carry out the work and prepare future generations of curriculum developers, assessment designers, and learning researchers to work in a more collaborative fashion with educational systems in order to develop more widely usable materials that fit into a coherent system of instructional guidance for teachers.

Measurement specialists have a role to play in developing coherent curriculum, instruction, and assessment systems, especially in helping subject-matter experts draw linkages between qualitative interpretations of student thinking and quantitative summaries needed to report achievement to external audiences and to enable valid comparisons beyond the classroom.

Notes

[1] Reproduced with permission from the Charles A. Dana Center, University of Texas at Austin and the Mathematics Assessment Resource Service, http:www.insidemathematics.org/performance-assessment-tasks.

[2] Marzano (2010) following from Wiggins (1996) distinguished between *standards-referenced* systems (whereby student achievement is reported in relation to performance standards) and *standards-based* systems (where again achievement is reported in relation to performance standards but a student is not permitted to move to the next level until competency at a given level has been demonstrated. Marzano also acknowledged that this distinction is not maintained in practice. Here we use the more familiar term, standards-based grading, without assuming that it implies grade retention or lockstep curriculum sequencing.

References

Bang, M., & Medin, D. (2010). Cultural processes in science education: Supporting the navigation of multiple epistemologies. *Science Education*, *94*, 1008–1026.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy, and Practices*, *18*(1), 5–25.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. Maidenhead, UK: Open University Press.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, *5*(1), 7–74.

Bowers, A. J. (2010). Grades and graduation: A longitudinal risk perspective to identify student dropouts. *Journal of Educational Research*, *103*, 191–207.

Brookhart, S. M. (2011). Starting the conversation about grading. *Educational Leadership*, *69*(3), 10–14.

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., Stevens, M. T., & Welsh, M. E. (2016). A century of grading research: Meaning and value in the most common educational measure. *Review of Educational Research*, *86*, 803–848.

Butler, R., & Nisan, M. (1986). Effects of no feedback, task-related comments, and grades on intrinsic motivation and performance. *Journal of Educational Psychology*, *78*, 210–216.

Cobb, P. (2000). The importance of a situated view of learning to the design of reserch and instruction. In J. Boaler (Ed.), *Multiple perspectives on mathematics teaching and learning* (pp. 45–82). Stamford, CT: Ablex.

Cobb, P., & Jackson, K. (2012). Analyzing educational policies: A learning design perspective. *Journal of the Learning Sciences*, *21*, 487–521.

Coburn, C. E., Penuel, W. R., & Geil, K. (2013). *Research-practice partnerships at the district level: A new strategy for leveraging research for educational improvement*. Berkeley, CA and Boulder, CO: University of California and University of Colorado.

Cole, M. (2010). What's culture got to do with it? Educational research as a necessarily interdisciplinary enterprise. *Educational Researcher*, *39*(6), 461–470.

Corcoran, T. B., Mosher, F. A., & Rogat, A. (2009). *Learning progressions in science: An evidence-based approach to reform*. Philadelphia, PA: Consortium for Policy Research in Education.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana: University of Illinois Press.

David, J. L., & Shields, P. M. (2001). *When theory hits reality: Standards-based reform in urban districts. Final narrative report*. Arlington, VA: SRI International. ERIC Number: ED480210

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, *125*, 627–668.

diSessa, A. A., & Minstrell, J. (1998). Cultivating conceptual change with benchmark lessons. In J. G. Greeno & S. V. Goldman (Eds.), *Thinking practices in learning and teaching science and mathematics* (pp. 155–187). Mahwah, NJ: Lawrence Erlbaum.

Doudna, J. A., & Charpentier, E. (2014). The new frontier of genome engineering with CRISPR-Cas9. *Science*, *346*(6213), 1258096.

Eccles, J. (2009). Who am I and what am I going to do with my life? Personal and collective identities as motivators of action. *Educational Psychologist*, *44*(2), 78–89.

Engeström, Y. (1991). Non scolae sed vitae discimus: Toward overcoming the encapsulation of school learning. *Learning and Instruction*, *1*, 243–259.

Ford, M. (2008). "Grasp of practice" as a reasoning resource for inquiry and nature of science understanding. *Science and Education*, *17*(2), 147–177.

Foster, D., & Poppers, A. E. (2011). How can I get them to understand? In P. E. Noyce & D. T. Hickey (Eds.), *New frontiers in formative assessment* (pp. 13–67). Cambridge, MA: Harvard Education Press.

Gallagher, H. A., Arshan, N., & Woodworth, K. (2017). Impact of the National Writing Project's College-Ready Writers Program in high-need rural districts. *Journal of Research on Educational Effectiveness*, *10*, 570–595.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist*, *18*, 519–521.

Glaser, R. (1984). Education and thinking: The role of knowledge. *American Psychologist*, *39*(2), 93–104.

Gravemeijer, K. (1994). Educational development and developmental research in mathematics education. *Journal for Research in Mathematics Education*, *25*, 443–471.

Guskey, T. R. (2009). *Practical solutions for serious problems in standards-based grading*. Thousand Oaks, CA: Corwin Press.

Hart, B., & Risley, T. R. (1999). *The social world of children: Learning to talk*. Baltimore, MD: Brookes Publishing.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112.

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

Hickey, D. T., & Pellegrino, J. W. (2005). Theory, level, and function: Three dimensions for understanding transfer and student assessment. In J. P. Mestre (Ed.), *Transfer of learning from a modern multidisciplinary perspective* (pp. 251–293). Charlotte, NC: Information Age.

Holland, D., & Lave, J. (2009). Social practice theory and the historical production of persons. *Actio: An International Journal of Human Activity Theory*, *2*, 1–15.

Honig, M. I., & Hatch, T. C. (2004). Crafting coherence: How schools strategically manage multiple, external demands. *Educational Researcher*, *33*(8), 16–30.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*, 254–284.

Lave, J. (1993a). Learning in practice. In S. Chaiklin & J. Lave (Eds.), *Understanding practice: Perspectives on activity and context* (pp. 3–32). New York, NY: Cambridge University Press.

Lave, J. (1993b). Situating learning in communities of practice. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 17–36). Washington, DC: American Psychological Association.

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

Lee, C. D. (1995). A culturally based cognitive apprenticeship: Teaching African American high school students skills in literary interpretation. *Reading Research Quarterly*, *30*, 608–630.

Lehrer, R., & Schauble, L. (2015). Learning progressions: The whole world is NOT a stage. *Science Education*, *99*, 432–437.

Marion, S., & Shepard, L. (2010). *Let's not forget about opportunity to learn: Curricular support for innovative assessments*. Dover, NH: National Center for the Improvement of Educational Assessment, Center for Assessment. Available at https://www.nciea.org/sites/default/files/publications/Marion_Shepard_Curricular_units_042610.pdf

Marzano, R. J. (2010). *Formative assessment and standards-based grading*. Bloomington, IN: Marzano Research Laboratory.

McMillan, J. H. (2001). Secondary teachers' classroom assessment and grading practices. *Educational Measurement: Issues and Practice*, *20*(1), 20–32.

Moll, L. C., Amanti, C., Neff, D., & Gonzalez, N. (1992). Funds of knowledge for teaching: Using a qualitative approach to connect homes and classrooms. *Theory Into Practice*, *31*(2), 132–141.

Nasir, N. S. (2000). "Points ain't everything": Emergent goals and average and percent understandings in the play of basketball among African American students. *Anthropology and Education Quarterly*, *31*(3), 283–305.

Nasir, N. S., Rosebery, A., Warren, B., & Lee, C. D. (2014). Learning as a cultural process: Achieving equity through diversity. In R. K. Sawyer (Ed.), *Cambridge handbook of the learning sciences* (2nd ed., pp. 686–706). New York, NY: Cambridge University Press.

National Research Council. (2000). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academies Press.

National Research Council. (2001). *Knowing what students know*. Washington, DC: National Academies Press.

National Research Council. (2003). *Assessment in support of instruction and learning: Bridging the gap between large-scale and classroom assessment*. Washington, DC: National Academies Press.

National Research Council. (2012a). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press.

National Research Council. (2012b). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Washington, DC: National Academies Press.

National Research Council. (2014). *Developing assessments for the Next Generation Science Standards*. Washington, DC: National Academies Press.

NGSS Lead States. (2013). *Next Generation Science Standards: For states, by states*. Washington, DC: The National Academies Press.

O'Connor, K. (2002). *How to grade for learning: Linking grades to standards*. Thousand Oaks, CA: Corwin Press.

Pearson. (2005). Achieving student progress with scientifically based formative assessment: A white paper from Pearson. No longer available at www.pearsoned.com

Pellegrino, J. W. (2014). A learning sciences perspective on the design and use of assessment in education. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 233–252). New York, NY: Cambridge University Press.

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist*, *51*(1), 59–81.

Penuel, W. R. (2015). Learning progressions as evolving tools in joint enterprises for educational improvement. *Measurement: Interdisciplinary Research and Perspectives*, *13*(2), 123–127.

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

Penuel, W. R., & Shepard, L. A. (2016b). Social models of learning and assessment. In A. A. Rupp & J. P. Leighton (Eds.), *Handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 146–173). Hoboken, NJ: John Wiley.

Penuel, W. R., Van Horne, K., Severance, S., Quigley, D., & Sumner, T. (2016). Students' responses to curricular activities as indicator of coherence in project-based science. In C.-K. Looi, J. L. Polman, U. Cress, & P. Reimann (Eds.), *Proceedings of the 12th International Conference of the Learning Sciences* (Vol. 2, pp. 855–858). Singapore: International Society of the Learning Sciences.

Popham, W. J. (2008). *Transformative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Reiser, B. J. (2014, April). *Designing coherent storylines aligned with NGSS for the K-12 classroom*. Paper presented at the National Science Education Leadership Association Meeting, Boston, MA.

Rogoff, B. (2003). *The cultural nature of human development*. New York, NY: Oxford University Press.

Rogoff, B., Baker-Sennett, J., Lacasa, P., & Goldsmith, D. (1995). Development through participation in sociocultural activity. In J. Goodnow, P. Miller, & F. Kessel (Eds.), *Cultural practices as contexts for development* (pp. 45–65). San Francisco, CA: Jossey-Bass.

Rogoff, B., Moore, L., Najafi, B., Dexter, A., Correa-Chavez, M., & Solis, J. (2007). Children's development of cultural repertoires through participation in everyday routines and practices. In J. E. Grusec & P. D. Hastings (Eds.), *Handbook of socialization: Theory and research* (pp. 490–515). New York, NY: Guilford Press.

Rogoff, B., Paradise, R., Arauz, R. M., Correa-Chavez, M., & Angelillo, C. (2003). Firsthand learning through intent participation. *Annual Review of Psychology*, *54*, 175–203.

Roschelle, J., Knudsen, J., & Hegedus, S. J. (2010). From new technological infrastructures to curricular activity systems: Advanced designs for teaching and learning. In M. J. Jacobson & P. Reimann (Eds.), *Designs for learning environments of the future: International perspectives from the learning sciences* (pp. 233–262). New York, NY: Springer.

Ruiz-Primo, M. A., Li, M., Wills, K., Giamellaro, M., L, M. C., Mason, H., & Sands, D. (2012). Developing and evaluating instructionally sensitive assessments in science. *Journal of Research in Science Teaching*, *49*, 691–712.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching*, *39*, 369–393.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *17*(2), 119–144.

Saxe, G. B. (1988). The mathematics of child street vendors. *Child Development*, *59*, 1415–1425.

Schmidt, W. H., Jakwerth, P. M., & McKnight, C. C. (1998). Curriculum sensitive assessment: Content *does* make a difference. *International Journal of Educational Research*, *29*, 503–527.

Shepard, L. A. (1997). *Measuring achievement: What does it mean to test for robust understanding?* Princeton, NJ: Educational Testing Service.

Shepard, L. A., Davidson, K. L., & Bowman, R. (2011). *How middle school mathematics teachers use interim and benchmark assessment data*. CSE Technical Report 807. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Shepard, L., Hammerness, K., Darling-Hammond, L., & Rust, F. (2005). Assessment. In L. Darling-Hammond & J. Bransford (Eds.), *Preparing teachers for a changing world* (pp. 275–326). San Francisco, CA: Jossey-Bass.

Shepard, L. A., Penuel, W. R., & Davidson, K. L. (2017). Design principles for new systems of assessment. *Phi Delta Kappan*, *98*(6), 47–52.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189.

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic molecular theory. *Measurement: Interdisciplinary Research and Perspective*, *4*(1–2), 1–98.

Soep, E. (2006). Critique: Assessment and the production of learning. *Teachers College Record*, *108*, 748–777.

Supovitz, J. A. (2006). *The case for district-based reform: Leading, building, and sustaining school improvement*. Cambridge, MA: Harvard Education Press.

Tzou, C. T., & Bell, P. (2010). *Micros and Me:* Leveraging home and community practices in formal science instruction. In K. Gomez, L. Lyons, & J. Radinsky (Eds.), *Proceedings of the 9th International Conference of the Learning Sciences* (pp. 1135–1143). Chicago, IL: International Society of the Learning Sciences.

Wang, M.T., & Eccles, J. S. (2013). School context, achievement motivation, and academic engagement: A longitudinal study of school engagement using a multidimensional perspective. *Learning and Instruction*, *28*, 12–23.

Wearne, D. & Hiebert, J. (1988). A cognitive approach to meaningful mathematics instruction: Testing a local theory using decimal numbers. *Journal for Research in Mathematics Education*, *19*(5), 371–384.

Wiggins, G. (1996). Honesty and fairness: Toward better grading and reporting. In T. R. Guskey (Ed.), *ASCD yearbook, 1996: Communicating student learning* (pp. 141–177). Alexandria, VA: Association for Supervision and Curriculum Development.

Wiser, M., Smith, C. L., & Doubler, S. (2012). Learning progressions as tools for curriculum development: Lessons from the Inquiry Project. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science* (pp. 359–403). Rotterdam, The Netherlands: Sense Publishers.