Using Formative Assessment to Create Coherent and Equitable Assessment Systems

Lorrie A. Shepard, William R. Penuel, & Kristen Davidson
University of Colorado Boulder

Most educators know that formative assessment is a powerful means to improve student learning. In the past, however, innovations intended to improve formative assessment practices have been derailed by accountability pressures resulting from high-stakes testing requirements. The Every Student Succeeds Act (ESSA), enacted in December 2015, carries forward many of the same testing requirements as No Child Left Behind (NCLB), but it removes the most severe consequences and returns greater control to states and school districts. Thus, there is hopeful talk among education leaders that a better balance might be possible between formative and summative assessment purposes. In particular, ESSA provides funding for states to develop new forms of assessments, to improve the alignment of assessments with curricula and instructional materials, and to develop more complete, balanced assessment systems that bring together summative, interim, and formative assessments for various purposes.

Given that states and districts will want to develop systems that are responsive to local contexts, what guiding principles might there be to ensure the design of effective assessment systems that genuinely improve learning? Informed by the research evidence and practical lessons that have been learned from three decades of assessment reform, we propose that development of localized assessment systems be based in three principles: coherence, learning theory, and equity.

Building Coherent Assessment Systems

First, assessment systems must be built to ensure *coherence* with respect to the learning goals that are assessed – both *vertical coherence* among assessments at the classroom, school, district, and state levels, and *horizontal coherence* whereby assessments are aligned with curriculum, instruction, and professional learning. The idea of building a coherent system of assessments "from classroom to state" was first advanced in a National Research Council committee report, *Knowing What Students Know* (KWSK) (Pellegrino, Chudowsky, & Glaser, 2001, p. 9). Synthesizing findings from cognitive science on how people learn and innovations in measurement science, the report explained why the models of learning underlying both classroom and large-scale assessment had to be conceptually compatible. Both classroom and large-scale assessments should be based "on a well-developed model of learning" in such a way as to "signal worthy goals" (p. 248). This recommendation was in contrast to multiple-choice standardized test formats that have invited an undesirable coherence between classroom worksheets and the high-stakes tests they were forced to imitate.

Today's next-generation standards call for the integration of disciplinary core ideas and practices. Thus, these are the learning goals with which both formative and summative assessment tasks should be aligned. *KWSK* authors did not suggest, however, that the two levels of assessment needed to provide the same level of detail. Formative assessment tasks

should elicit meaningful demonstrations of students' developing expertise in a content area during on-going instruction and should be accompanied by the kinds of questions and activities needed to support next steps. Accountability assessment tasks should be designed to be substantively congruent with the same kinds of tasks and activities that are used instructionally, but they do not need to be an exhaustive set of all such tasks. Thus, a student would see on the culminating test or assignment the same kinds of problems with the same expectations for quality performance as had been practiced and improved upon throughout the year.

If assessments of next-generation standards are to play a role in improving student outcomes, then people at all levels of the system need a shared understanding of learning goals, or what in *KWKS* is referred to as vertical coherence. But a shared understanding of new standards requires learning opportunities for education leaders, teachers, parents, and community members to develop understanding of what is new about the learning goals. And just having a shared vision will not be enough: adults at all levels of the system need knowledge about *how* to improve in relation to those goals, and students will need to understand the aims of their own learning. Teachers especially need opportunities to provide input and to collaborate with colleagues, as well as resources, good models, and consistent messages about implementing "next generation standards," integrating disciplinary core ideas and practices, and practicing culturally inclusive instruction. Overall, if the system is to be coherent, solutions should be developed jointly among teachers, school leaders, school communities, and policymakers.

Vertical and horizontal coherence are important in any system so that all of the components and actors in the system are working toward the same goals instead of at cross-purposes. Indeed, this idea has been at the center of a great deal of research on learning in the disciplines. Unfortunately, however, in the past two decades, the amount of testing required by accountability mandates has precluded sustained advances in creating more coherent operational assessment systems. Instead, formative and summative assessments have been taken up disparately, often without recognizing which approaches have a sound research base and which do not. In the next section, we offer a condensed summary of the major formative assessment approaches (Penuel & Shepard, 2016). Our focus is on the theory of learning that grounds each approach, that is, on each model's assumptions and knowledge base for how student outcomes can improve.

Basing Assessment Systems in a Model of Learning

In addition to the *KWSK* requirement that assessment system coherence should be based on a shared *model of learning*, we have argued that the effectiveness of any given formative assessment intervention depends on the *adequacy* of its underlying theory of learning. Adequacy may be judged by the value of learning goals as well as the sufficiency of evidence demonstrating means for reaching those goals, attention of the model to motivation, participation, and identity as well as cognitive goals, and consequences for diversity and equity.

In their original, famous review on formative assessment, Black and Wiliam (1998) reported on distinct literatures. They called for theory building, but they did not offer an

integrative theory about how or whether various strategies from self-assessment, mastery learning, feedback, and motivation could be fit into a coherent whole. In theoretical frameworks and meta-analyses since that time, many authors have cited references on formative assessment without recognizing when those references offered incompatible views of learning goals and learning. Here we describe the differences among four approaches that have been promoted as formative assessment and explain why the explicit learning theories in the latter two categories hold greater promise for supporting more ambitious and equitable, next-generation visions of teaching and learning.

Data-driven decision-making. Data-driven decision making (DDDM) calls for the use of data from interim or benchmark assessments (sometimes marketed as "formative assessments") closely aligned to state standardized tests in order to identify areas in need of improvement. Promulgated in response to NCLB, the DDDM movement in education is most accurately portrayed as a policy theory of action. It does not have a theory of learning but rather derives from theories of organizational change (Deming, 1986; Senge, 1990). The DDDM theory of action holds that educators will set goals aligned with standards and standardized tests, examine data to evaluate progress toward those goals, determine the causes of results, implement new strategies to address areas of weakness, and continue to monitor effectiveness. Importantly, it is assumed that teachers will know how to remedy the problems identified by data or will seek additional training—an assumption that has been disconfirmed in low-performing schools that often lack this demanding level of capacity (Elmore, 2003).

Most studies examining DDDM focus on the convening of data teams and other data use processes, but there is little research on subsequent improvements of teaching or learning. At its best, DDDM encourages focused conversations in schools and districts around learning outcomes and a shared goal of continuous improvement. Limitations of DDDM arise because of the impoverished representation of learning goals offered by the majority of computerized tools. Typically, interim assessments help teachers know which students are the most in need of help and which objectives are most in need of reteaching (Shepard, Davidson, & Bowman, 2011). But they provide no substantive help about student thinking or how to intervene. In some cases, DDDM systems also foster an extrinsic view of motivation based on rewards and punishments. In these instances, because of proficiency results posted in the hallways or the nature of score reports, students know that they need to get, say, three more items right to reach proficiency. Setting goals by counting items, however, does not help students and teachers achieve proficiency. Students need opportunities to engage with rich curriculum materials, and teachers need opportunities to develop expertise so as to be able to use those materials to address the particular difficulties students are facing.

Strategy-focused formative assessment. Strategy-focused formative assessment approaches provide teachers with tools and practices they can use as part of classroom routines to provide feedback to students and engage students more actively in their own learning. Teachers learn how to use questioning techniques to elicit student thinking, prompt classroom talk that provokes revision of ideas, provide qualitative feedback to students on how to improve, and engage students in self- and peer-assessment. Although the strategies

typically associated with this approach are compatible with constructivist and sociocognitive theories of learning, they did not arise from these learning theories. Rather, these strategies were taken up based on empirical evidence showing the effectiveness of those particular instructional practices.

To be successful, strategy-focused approaches call for professional learning opportunities in which teachers can try out various techniques while fostering classroom environments in which students assume a more active role in their own learning. One highly visible example, the King's-Medway-Oxfordshire Formative Assessment Project (KMOFAP) (Black, Harrison, Lee, Marshall, & Wiliam, 2003), showed promise for changing teachers' assessment practices. In KMOFAP, Black and Wiliam offered strategies such as questioning techniques, providing specific feedback, making criteria explicit, and encouraging peer- and self-assessment. In addition, they designed the project to draw from research documenting the benefits of eliciting and building upon students' ideas, cultivating intrinsic motivation, and encouraging students' regulation of their own learning.

Strategy-focused approaches to formative assessment, however, remain agnostic as to the nature of learning goals, which might in fact be quite traditional. Ultimately, the guidance provided by a generic strategy-focused program may not be sufficient to support deep learning, because the assessment strategies are not tied to particular learning goals or evidence of progress within instructional activities. Little attention is paid to how becoming proficient involves different learning processes, depending on the content area, or how demonstrating proficiency in the subject matter requires assessment tasks that engage students in disciplinary practices.

Sociocognitive formative assessment. Sociocognitive formative assessment projects attend to the social nature of cognition and provide resources to assess students' understandings and skills as they participate in increasingly sophisticated practices common to disciplinary experts. The design of these interventions is intentionally grounded in empirically-supported "local instructional theories" of learning, according to which a sequence of instructional activities are devised to support students in developing proficiency (Gravemeijer, 2004). Instructional sequences are typically based either on a "learning progressions" (or "trajectories") approach that lays out empirically supported claims about how students' understanding develops toward specific disciplinary goals (Simon, 1995; Smith, Wiser, Anderson, & Krajcik, 2006), or on a "knowledge-in-pieces" (or "facets") view that presumes students' conceptions as less ordered or stable and more tied to students' experiences and the contours of particular problems (diSessa, 1988). In both cases, assessment materials are grounded in the disciplinary content area and provide substantive insights regarding typical difficulties and productive ways forward.

In addition to mastering content knowledge, sociocognitive interventions aim for students to participate in practices that develop expertise in a discipline, which can involve the development of students' dispositions and identities in accordance with the field. Thus, assessments attend to the development of students' thinking and reasoning moves toward that of disciplinary experts. They similarly employ multiple strategies, such as collaborative inquiry, expertly facilitated questioning and discussion, and qualitative feedback. These

instructional contexts for eliciting, interpreting, and responding to students' thinking are each guided and informed by learning model frameworks.

An example with evidence of positive learning outcomes is the Inquiry Project, which focuses on improving students' understanding about the nature of matter along a hypothetical learning progression that models the development of expert understanding (Smith et al., 2006). Another example is the Contingent Pedagogies project (Penuel, DeBarger, Boscardin, Moorthy, Beauvineau, Kennedy, & Allison, in press) designed to help teachers elicit and build upon students' ideas and experiences to develop their understanding of disciplinary core ideas in earth science in the context of constructing explanations and building models of phenomena. The interactive assessments designed for Contingent Pedagogies were linked to the *Investigating Earth Systems* curriculum materials and were intended to provide more fine-grained analysis of student thinking over two to three days of instruction.

In the Penuel and Shepard (2016) review, we noted the importance of both a well-articulated learning theory and discipline-specific learning goals as strengths of sociocognitive formative assessment projects. In contrast to quantitative displays of correct and incorrect answers, sociocognitive assessments offer qualitative accounts of students' reasoning and problem-solving thus providing insights that can better inform instruction. The investments required to develop such projects, however, has resulted in a lack of availability in many schools and districts. We also note that the potential of these interventions to advance learning equitably is limited to the extent that youth's own interests, experiences, and agency in setting learning goals are not foregrounded in teaching and assessment.

Sociocultural formative assessment. Sociocultural interventions share with the sociocognitive approach many of the same assumptions about the social nature of learning and development as well as a value for participation in disciplinary ways of knowing and doing. The two theories of learning diverge the most, however, in terms of how they respond to diversity in students' entry points and existing knowledge. Sociocultural interventions take much more of a transformative stance, which allows for diverse pathways through disciplinary core ideas and practices that build upon familial and community practices.

Sociocultural models of learning and assessment recognize that students bring to the learning environment important knowledge, interests, and experiences from their daily lives that should inform curriculum and instruction. The aim of teaching from this perspective is to help students navigate between emerging fluency with disciplinary ways of knowing, doing, and being and the ways of knowing, doing, and being that are valued in their own communities (Bang & Medin, 2010). A key purpose for assessment is to elicit and make use of students' experiences and interests to inform the course of instruction and to help set goals for learning.

An example with promising evidence of effectiveness is the Bellevue-University of Washington (UW) Curriculum Redesign Partnership, which repurposes multiple units of study from the district's elementary science curriculum in ways that expand students' agency in the classroom and leverage the diversity of students' interests relevant to the focal topics of the units. One strand of the redesign uses a "challenge-based learning cycle" (Schwartz, Lin, Brophy, & Bransford, 1999), in which teachers first present an overarching challenge intended

to launch a cycle of inquiry, followed by elicitation of students' initial ideas regarding the challenge, teacher-led and student-designed investigation, revision of ideas, and a public presentation by students of their conclusions about the challenge. A second strand draws on strategies such as photo-elicitation to document everyday lives of people in communities (Clark-Ibañez, 2004). At the beginning of a unit on microbes and health, for example, students take photos of things or activities they do in daily life to prevent disease and stay healthy. They then share these photos in class, as a way to bring personally relevant experiences into the classroom to launch the unit. Their documentation also helps shape a student-led investigation focused on students' own questions, which are refined as students encounter key ideas in microbiology.

The kinds of assessment systems within which sociocultural interventions fit and help to support are ones that, for the most part, are yet to be developed. To succeed in today's educational systems, sociocultural interventions like the one implemented in Bellevue need significant support from stakeholders. Most specifically a local curriculum that sought to make productive use of students' interests, experiences, and funds of knowledge would have the greatest chance for success if metrics used for accountability purposes documented, valued, and mirrored disciplinary learning goals pursued in such projects.

Learning from the Past and Designing for the Future

States and districts seeking to design new assessment systems have the opportunity to build coherent systems in which daily, informal, instructionally-grounded formative assessments are based on the same model of learning as the assessment used to gather accountability evidence.

Painful lessons from the past, however, remind us that creating a coherent and effective assessment system does not mean building one assessment instrument to serve both formative and accountability purposes. "Tests worth teaching to" did not work well in the 1990s because even the best high-stakes assessments had to conform to standardization requirements. And even if our large-scale tests send good signals about worthy goals, tests without local guidance about how to improve performance do not automatically lead to better student achievement.

A very different approach would be *not* to start with building accountability tests as the primary instrument of educational reform but to begin instead with curriculum and instructional practices. Sociocognitive and sociocultural theories of learning offer the greatest promise for improving teaching and learning and to advance equitable learning opportunities because they work to particularize the meaning of ambitious learning goals in respective disciplines and they attend to the supports necessary to help students reach those goals. Learning progressions and facets are two sophisticated ways to create coherent assessment systems based on a shared model of learning, but they are by no means the only ways. Both the Contingent Pedagogies and Bellevue projects illustrate how reasonably good, locally-selected curricula can be further deepened and enhanced by co-designing formative and end-of-unit assessments connected to those curricula. The National Writing Project is another example with a long history where it is possible to maintain—in unit assessments and

program evaluation—the same features of student writing that are valued and attended to in on-going instruction.

Creating assessment systems that genuinely support learning requires the redesign of both classroom and large-scale assessments. Ideally, system design would start with curriculum and instruction. Then, portfolios, extended projects, as recommended to assess Next Generation Science Standards (Pellegrino, Wilson, Koenig, & Beatty, 2014), and perhaps common assessment tasks as being piloted in New Hampshire could be devised in ways that would best support teaching and learning. Only after building horizontal coherence would designers ask how common scoring rubrics and scoring trainings could make it possible to score *some* projects for both classroom grades and accountability reporting. Even such ground-up designs will need to keep an eye on protecting the authenticity of instructional processes and balancing workload burden versus the professional development benefits of comparable scoring for accountability purposes. Because the strength of his approach comes from embedding assessments in curriculum and shared instructional practices, it is unlikely to be feasible when designers must allow for very different curricula.

Of course, states and districts often are required to start with summative accountability assessments without the substantive supports of a shared curriculum. This is more difficult, but the same principles apply. To advance deep learning and equity, states should design assessments based on valid learning theories and should also consider safeguards against the kinds of distortions that have been caused by previous top-down mandates. They should especially avoid reliance on multiple-choice-only formats that correlate with learning goals but lead to distorted coherence between classroom learning tasks and assessments.

Well-developed theories of learning not only allow a state assessment to "signal worthy goals" but also provide evidence-based pathways for attaining those goals along with examples of how students' personal interests and community practices can be connected to disciplinary learning. At the large-scale level, full representation of core ideas and practices is essential to forestall teaching to a narrow subset of goals. At the same time, to safeguard against the burden of excessive testing times, it would be possible to rotate across years among specific intersections of core ideas and practices.

When states or districts also seek to play a role at the classroom level, attending to the more particularized models of learning by which embedded formative assessments are connected to ongoing instruction becomes more difficult, if they serve jurisdictions with multiple curricula. In such circumstances, a limited number of "replacement units" could be co-designed with local districts and made available on a voluntary basis as powerful examples to model learning of selected core ideas and practices. Replacement units are curricular units of study with lesson plans, instructional activities, embedded formative assessments, and end-of-unit summative assessments coherently designed to enable implementation of new instructional approaches. In this context of state initiatives, replacement units could also be designed to exemplify a coherent relationship between formative questions and tasks that provide qualitative insights and feedback leading to increasing proficiency on some of the same learning targets intended for accountability assessments.

Whether built by starting from curriculum and instruction or designed of necessity primarily for accountability purposes, full consideration should be given to the ways that

official assessments drive the discourse about what counts in teaching and learning. They shape what students, families, educators, community members, and policymakers understand and value about learning. Instead of merely ranking or categorizing schools, educators and assessment designers must instead focus on the substance of learning. They might consider, for example, publishing student essays or science investigations representing a range of students' work. If families choose to opt their students out of taking state assessments, would they elect instead to have their student receive feedback on a common writing assignment or problem set? If ESSA does not allow matrix sampling to reduce burden and at the same time increase the reach of accountability assessments, would it be possible to pilot some combination of adaptive scales, anchor tasks and matrix sampling of other tasks? The main idea should be to design an assessment system true to learning-focused principles and also to experiment with and continue to improve that system.

References

Bang, M., & Medin, D. (2010). Cultural processes in science education: Supporting the navigation of multiple epistemologies. *Science Education*, *94*(6), 1008-1026.

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

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

Clark-Ibañez, M. (2004). Framing the social world with photo-elicitation interviews. The *Behavioral Scientist*, *47*(12), 1507-1527.

Deming, W. E. (1986). *Out of the crisis*. Cambridge, MA: MIT Press.

diSessa, A. A. (1988). Knowledge in pieces. In G. Forman & P. Pufall (Eds.), Constructivism in the computer age (Vol. 49-70). Hillsdale, NJ: Erlbaum.

Gravemeijer, K. (2004). Local instruction theories as means of support for teachers in reform mathematics education. *Mathematical Thinking and Learning*, *6*(2), 105-128.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know*. Washington, DC: National Academies Press.

Pellegrino, J. W., Wilson, M. R., Koenig, J. A., & Beatty, A. S., (Eds.) (2014). *Developing assessments for the next generation science standards*. Washington, DC: National Academies Press.

Penuel, W. R., DeBarger, A. H., Boscardin, C. K., Moorthy, S., Beauvineau, Y., Kennedy, C., Allison, K. (in press). Investigating science curriculum adaptation as a strategy to improve teaching and learning. *Science Education*.

Penuel, W. R., & Shepard, L. A. (2016). Assessment and teaching. In D. H. Gitomer & C. A. Bell (Eds.), *Handbook of research on teaching*. Washington, DC: AERA.

Schwartz, D. L., Lin, X., Brophy, S., & Bransford, J. D. (1999). Toward the development of flexibly adaptive instructional designs. In C. Reigeluth (Ed.), *Instructional design theories and models: A new paradigm of instructional theory* (Vol. II, pp. 183-214). Mahwah, NJ: Earlbaum.

Senge, P. (1990). *The fifth discipline: The art and practice of the learning organization*. New York: Doubleday.

Shepard, L. A., Davidson, K. L., & Bowman, R. (2011). How middle school mathematicss 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).

Simon, M. A. (1995). Reconstructing mathematics pedagogy from a constructivist perspective. *Journal for Research in Mathematics Education*, *26*, 114-145.

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 & Perspective*, *4*(1&2), 1-98.