Re-Balancing Assessment:

Placing Formative and Performance Assessment at the Heart of Learning and Accountability

By Peter Hofman, Bryan Goodwin, and Stuart Kahl

McREL INTERNATIONAL

measured progress™

McREL International
4601 DTC Blvd., Suite 500, Denver, CO 80237 USA
Phone: 303.337.0990 • Fax: 303.337.3005
Website: www.mcrel.org • E-mail: info@mcrel.org

About McREL

McREL International is a nonprofit, nonpartisan organization dedicated to improving education outcomes for all students through applied research, product development, and professional service to teachers and education leaders across the U.S., in Australia, and in the Pacific Region.

Measured Progress
100 Education Way, P.O. Box 1217, Dover, NH 03821 USA
Phone: 603.749.9102 • Fax: 603.749.6398
Website: www.measuredprogress.org
E-mail: mail@measuredprogress.org

About Measured Progress

Measured Progress is a not-for-profit company specializing in a wide range of assessments, from effective classroom assessment practices to high-stakes, large-scale assessments for general and special student populations. We also work with state and local educators to build their capacity in assessment literacy.

Hofman, P., Goodwin, B., & Kahl, S. (2015). *Re-balancing assessment: Placing formative and performance assessment at the heart of learning and accountability*. Denver, CO: McREL International.

Acknowledgements

The authors thank David Hopkins, Michele McLaughlin, Ray Pecheone, Tim Waters, and Robin Wisniewski for their review and feedback on drafts of this paper, which aided greatly in its development. Thanks also to Collaborative Communications, Roger Fiedler, Heather Hein, and Colleen Ingham for their editing and production work.

20150218

Re-Balancing Assessment:

Placing Formative and Performance Assessment at the Heart of Learning and Accountability

By Peter Hofman, Bryan Goodwin, and Stuart Kahl

Overview

What gets tested is what gets taught. Ever since Frederick Kelly, Dean of the College of Education at the University of Kansas, introduced the multiple-choice test in 1914, we've increasingly tested what's easy to measure, not necessarily what matters. While a fine approach for many basic skills, it falls far short of facilitating the deeper learning demanded in the 21st century. When high stakes are added to these tests, they further narrow the focus of teaching and learning, place unproductive stress on educators, and diminish student engagement and motivation.

These days, a growing chorus of parents, educators, and policymakers is voicing frustration and anger with top-down accountability and high-stakes testing. As members of two not-for-profit education organizations—one focused on assessment and the other on research and instructional practices—we find nothing wrong with testing itself; indeed, we believe evidence of what students know and can do should be at the heart of schooling. We are concerned, however, about what seems to be an almost myopic focus on high-stakes accountability based on tests of basic knowledge and skills to drive improvements in educational outcomes, to the exclusion of using formative and performance assessment to facilitate student growth and deeper learning (i.e., the ability to apply knowledge and skills in novel situations) (Pelligrino & Hilton, 2012).

In keeping with the opinions of the Gordon Commission on the Future of Assessment in Education, we, too, believe that high-stakes testing should not be "the only—or even the major—tool for improving student outcomes" (Gordon Commission, 2013, p. 7). We do not underestimate the complex challenges involved in improving public education to meet the demands of the 21st century. However, solving these complex challenges will require employing a different set of drivers—ones that reflect what Daniel Pink popularized as "Motivation 3.0" in his book, *Drive: The Surprising Truth About What Motivates Us* (2009): namely, autonomy (some choice in how we solve a problem), mastery (a sense of competence in what we're doing), and purpose (understanding why what we're doing is important).

Thus, we call for replacing our current, unbalanced formula of summative assessments and external pressure with a new, more balanced formula for assessment and accountability centered around ***curriculum-embedded performance assessments*** or ***CEPAs***—classroom-based instructional units that provide multiple opportunities for learning and both formative and summative evidence-gathering. Done well, CEPAs can harness the power of real-time feedback, personalized learning, and real-world application to help students develop requisite foundational knowledge *and* deeper learning. Moreover, as we will discuss, CEPAs may be used for summative, including state accountability, purposes. While they are not a silver bullet, CEPAs could drive many other beneficial changes in the education system, including better classroom practice, more motivating and engaging school environments, and greater professional collaboration among educators. Perhaps most promising, CEPAs could undergird a state-level accountability system that measures what matters most: the extent to which students are developing and demonstrating the kinds of deeper learning they will need for success in college, career, and life.

Outgrowing Once-Promising Formulas

F=ma. This simple formula for relating the force exerted on an object, the mass of the object, and its acceleration, along with several other laws, comprise a body of science known as Newtonian physics—elegant formulas and laws that for centuries appeared to capture and predict the world as we knew it.

There was just one problem. By the 20th century, scientists observed that on the outer edges of science, Newton's tidy laws no longer seemed to work. At a grand scale, they were insufficient to predict the behavior of large bodies. And at the subatomic scale, photons and electrons seemed to defy the rules of Newtonian physics.

We have since begun to unravel these mysteries with a new set of rules, including theories of relativity and quantum mechanics.

So it may be with education reform.

Measuring What's Easy to Measure

For more than two decades, our nation has labored under something like a Newtonian notion of what it takes to improve student performance. We assume if we can manage to determine the right formula for high-stakes testing and accountability, we can make our school systems act more predictably, turning seeming inertia into an object in motion. We might sum up this formula as the following:

It is a straightforward formula that seems sensible enough. However, its formulation coincided with the growth of large-scale statewide testing programs that require student responses to be easily, efficiently, and "objectively" scored. These testing programs rely on standardized tests comprised almost exclusively of selected-response items that overemphasize basic skills and factual recall. These types of tests tend to drive correspondingly low-level instructional content and practices. Certainly, basic skills and knowledge are important building blocks of learning, but are hardly what anyone would envision as the sole desired outcome of schooling. Yet, to date, our formula for driving reform, as reflected in many bipartisan federal and state policies, has primarily relied on measuring student learning with large-scale, low-level standardized assessments. This is our first concern. In effect, schools and teachers are held accountable for achieving what amounts to only factual recall and basic skills.

Our second concern is that the higher stakes attached to these standardized tests have led to a proliferation of even more standardized "interim" tests to predict student results on the all-important end-of-year assessments. Students now find themselves spending increasing amounts of time taking tests instead of learning. A recent survey by the Council of Chief State School Officers (CCSSO) determined that between preschool and their senior year of high school, U.S. students take an average of 113 standardized tests (Kamenetz, 2014). A study of 14 school districts conducted by the Center for American Progress found students in grades 3–8 taking an average of 10 and as many as 20 standardized tests per year (Lazarín, 2014). Recently, thousands of high school seniors in three Colorado school districts appeared to reach their breaking point when they walked out of state-mandated science and social studies exams. "We have grown up taking standardized testing—since third grade," one student told the *Denver Post*. "This particular protest comes as a result of this frustration [of] taking these tests we don't feel are adequate" (Gorski, 2014). In response to the growing chorus of concerns, CCSSO and the Council of Great City Schools (2014) issued a pledge to cut back on unnecessary testing for students.

We are not anti-testing. Far from it. We are, however, anti-*ineffective* testing—testing that solely focuses on basic skills and foundational knowledge, does not involve students documenting their thinking and the results of their work, does not inform instruction or learning, hinders rather than fosters student growth, and is meaningless to students and teachers—in short, testing that wastes precious time, resources, and energy. We know we are not alone. Much, if not most, of the opposition to testing is to "standardized testing," particularly in the form of selected-response tests and the associated time spent teaching test-taking tricks rather than engaging students in deeper learning.

Testing Our Way to a Performance Ceiling

In some ways, the rationale underlying the current formula for school accountability is reasonable. In business, few enterprises succeed without ambitious goals or performance data. We, too, believe standards can be important tools to raise expectations. And behavioral psychology tells us that people do respond to rewards and punishments. Within education, shining the bright light of accountability on schools has prompted some to make dramatic changes that improved student outcomes, delivering what David Hopkins, former education advisor to British Prime Minister Tony Blair, has described as a "short, sharp shock" that shakes systems "out of complacency" or helps them in "directing their attention to a limited number of measurable goals" (Hopkins, 2013, p. 9).

Here in the U.S., there is some evidence that No Child Left Behind's (NCLB) Adequate Yearly Progress and associated sanction provisions have been associated with a modest, positive impact on student test scores (Ahn & Vigdor, 2013, 2014). An analysis of trend data in 25 states since the implementation of NCLB found some correlations between high-stakes testing pressure, gains in student achievement, and narrowing of achievement gaps. However, it also showed that "students were progressing in math at a much faster rate before the national high-stakes testing movement spawned by NCLB" and that in mathematics, early gains related to testing pressure plateaued and then declined. In short, the researchers concluded, "these data suggest that pressure has diminishing returns for math achievement over time" (Nichols, Glass, & Berliner, 2012, p. 26).

Long-term trends from the National Assessment of Educational Progress (NAEP) reveal a similar pattern (Institute of Education Sciences, 2013). Students' scores have generally improved or remained flat since NAEP was first administered in 1971. Interestingly, during the NCLB era, only 13-year-olds showed improvement in reading and math; 9- and 17-year-olds did not. Digging more deeply into the data suggests that the past two decades of test-driven, high-stakes reform have had some positive impact in raising the performance of low-achieving students and ensuring more students can demonstrate basic, low-level reading and math skills. However, the early progress made in narrowing achievement gaps appears to have slowed. In many areas, as a nation, we appear to have hit a performance ceiling. As a result, we have been unable to help great numbers of students, especially older ones, master high-level reading and math skills.

Hopkins observed a similar phenomenon in the U.K. "The problem is that such top-down strategies have a very limited half-life," he wrote. "Once the school or system has begun to improve and to take ownership of its own development, the continuing pressure for external accountability becomes oppressive, alienating, and counter-productive" (Hopkins, 2013, p. 9). Across the U.K., for example, he observed that reading scores rose in response to external accountability pressures, but then leveled off as those pressures offered "little guidance as to how to create more productive, instructional, and curriculum pathways for students" or encourage "assessment for learning," which, research shows, when utilized well, can accelerate "student engagement, learning, and achievement" (Hopkins, 2013, p. 9).

This performance ceiling may be most evident in international comparisons of student performance. Since the advent of high-stakes testing in the U.S., American student performance on international comparisons has declined in relative terms, largely as our own incremental gains in performance have been eclipsed by other nations that have made more substantive changes to their education systems and progress in student outcomes. For example, on the 2012 Program for International Student Assessment (PISA), which attempts to measure higher-order skills, U.S. students performed below average in math and about average in reading and science, with little change over time. In fact, the U.S. had a higher proportion than average of the lowest-performing students and a lower proportion of top-performing ones (Office for Economic Cooperation and Development [OECD], 2012).

While we in the U.S. have been using high-stakes testing to drive system improvements, leading performers such as the city of Shanghai, Singapore, and Finland dramatically changed their focus from *teaching facts* to *deeper learning*, from narrowly focused curricula to providing students with some autonomy and personalized learning choices, and away from high-stakes test performance as the sole goal of education to the development of well-rounded graduates with highly toned cognitive and non-cognitive skills and intelligences (OECD, 2012).

Confronting Unintended Consequences

In addition to failing to deliver the desired results, our current system of high-stakes testing and accountability appears to have had many unintended and counter-productive consequences:

- **Increasing stress levels for professionals.** A recent MetLife Foundation study found that half of teachers (51%) and principals (48%) were under great stress at least several days per week, significantly higher percentages than found in similar studies carried out in previous decades (MetLife, 2012). Such stress, it turns out, impedes performance. As noted by Po Bronson and Ashley Merryman in their book, *Top Dog* (2013), this kind of stress can inhibit performance, adversely affecting decision-making and our ability to take action. On high alert to avoid mistakes, we actually make more of them.
- **Little positive impact on classroom teaching.** University of Virginia researchers examined the classroom experiences of 994 students from across

the U.S. in grades 1, 3, and 5, and determined that, despite efforts to ensure greater teaching consistency through standards-based reform efforts, just *7 percent* received high-quality instruction and emotional support over all three years (Pianta, Belsky, Houts, & Morrison, 2007). While these results might not differ from any obtained prior to the passage of NCLB, they certainly illustrate that its implementation did not result in widespread improvement in instructional quality. In fact, some researchers have noted that our current reform efforts may actually be *reducing* the effectiveness of classroom instruction. At the advent of test-based reform efforts, researchers studying student engagement observed that teachers felt compelled to march through the curriculum, teaching what would be on "the test" with little attention to explaining its importance (Csikszentmihalyi, Rathunde, & Whalen, 1993).

- **Disengaged students.** Researchers have observed that test-harried teachers often have little time to pause and respond to teachable moments or episodes of spontaneous student curiosity, thus reducing student engagement and interest in school (Engel, 2011). Students themselves often find the high-stakes assessments to be meaningless. Research has found that a simple $10 incentive persuaded students to take high-stakes tests more seriously, resulting in a bump in performance that ostensibly reflected the equivalent of *six months* of additional learning (Levitt, List, Neckerman, & Sadoff, 2012). Preparing for what some students think is a meaningless test seems like a waste of time and effort. Why meaningless? Although some test results affect students' grades, promotions, and graduation, many do not.

Pivoting to a New Formula

Our point here is not that our current system of summative assessment, external pressure, and accountability is all bad, or that no good has come from the past 25 years of education reform. To the contrary, these policies have helped focus our entire system of education on using data, raising expectations for learning, and, in the case of No Child Left Behind, focusing on the moral imperative of helping *all* children succeed. However, we have concluded that we have reached the point of diminishing returns with this formula. It is now time to pivot to a new paradigm for reform. As with Newtonian physics, our current formula of top-down, high-stakes, and test-based accountability does not capture the best and most current knowledge about what works in education reform. While it has delivered some "shock treatment" to the system and laid important groundwork for change, continuing with this approach will neither support widespread shifts in classroom practice nor universally improve learning to levels needed for student success in the 21st century.

It is time for a new formula that starts with changing how we perceive and use assessment to drive change. The problem and the potential of better approaches to assessment were best summed up in a policy brief from the Gordon Commission on the Future of Assessment in Education, established by the Educational Testing Service in 2011:

> Throughout the long history of educational assessment in the United States, it has been seen by policymakers as a means of enforcing accountability for the performance of teachers and schools. But, as long as that remains their primary purpose, assessments will never fully realize their potential to guide and inform teaching and learning. … The problem is that other purposes of assessment, such as providing instructionally relevant feedback to teachers and students, get lost when the sole goal of states is to use them to obtain an estimate of how much students have learned in the course of a year. (Gordon Commission, 2013, p. 7)

Over the long term, the question becomes, how do we actually support teachers in helping students achieve the higher expectations that are today's norm? We believe the answer lies in re-balancing our use of assessment for teaching, learning, and accountability.

Driving Change with Formative and Performance Assessment

Although richer and more rigorous expectations are driving accountability assessments in the right direction, to date, they merely represent an incremental step. The assessments developed by the multi-state assessment consortia are more performance-based than many existing state tests, but far less so than initially proposed. Unfortunately, the same realities that spawned the proliferation of standardized testing over the past two decades (cost, testing time, and test security concerns) have remained, leaving high-stakes, on-demand, summative assessments comprised of a significant number of selected-response items as the dominant measure of student learning.

In short, we are still stuck.

Yet we know that a broad spectrum of individuals and entities are seeking better alternatives, with less emphasis on top-down, low-level, single test-based accountability and more emphasis on performance-based assessment. The time is right for a new formula centered on the game-changing assessment system we propose. Rather than being hatched in isolation, what we advocate reflects research, evidence, and experience, and builds on the solid foundation laid by others, including the National Research Council's (NRC)(2012) calls for greater emphasis in schooling on *deeper learning*—that is, the development and application of a combination of cognitive, intrapersonal, and interpersonal competencies. The NRC identified several research-based methods for supporting student mastery:

- **Encourage elaboration, questioning, and explanation**—for example, prompt students reading a history text to think about the author's intent and/or to explain specific information and arguments as they read, either silently to themselves or to others.
- **Engage learners in challenging tasks** while also supporting them with guidance, feedback, and encouragement to reflect on their own learning processes.
- **Prime student motivation** by connecting topics to students' personal lives and interests, engaging students in problem solving, and drawing attention to the knowledge and skills students are developing and their relevance, rather than grades or scores.
- **Use "formative" assessments [evidence-gathering techniques]**, which continuously monitor students' progress and provide feedback to teachers and students for use in adjusting their teaching and learning strategies. (pp. 9–10)

Central to these recommendations is a fundamental shift in our theory of action for reform. On the surface, it might seem that relying more on formative and performance assessment is merely a technical fix. In reality, formative and performance assessment represent a different way of thinking about how and why learning occurs—about the roles teachers and students play and how they spend their time.

Too often, in traditional learning environments, students learn because they are *compelled* to do so by external factors—as reflected in the age-old question, "Will this be on the test?" Formative and performance assessment begin at a different starting point—one that aims to drive intrinsic motivation for learning by asking and answering *why* students ought to learn something and making it relevant for them. At the same time, formative assessment aims to give students the feedback

PRINCIPLES TO GUIDE A NEW FORMULA FOR ASSESSMENT, ACCOUNTABILITY, & LEARNING

- At all levels, accountability and assessment should reflect both foundational knowledge/skills and deeper learning.
- Consistent expectations between state accountability and local assessment are required to help all students achieve deeper learning.
- Accountability systems must be upgraded to reflect and deliver on the reality that policymakers and educators at different governance levels have different data needs.
- Curriculum-embedded performance assessments—CEPAs for short—should lie at the center of the teaching and learning process and accountability.
- Performance tasks in on-demand components of state assessments and in CEPAs should provide high-quality measures of student learning.
- Results from locally scored (but state audited) summative tasks within CEPAs should play a role in state accountability assessment.

they need so that they *all* might be successful. In short, both formative and performance assessment go to the heart of where learning and teaching happen. Effective implementation will improve teaching practice and will engage and motivate students to take ownership of their own learning; it will enhance students' higher-order cognitive and non-cognitive skills, better preparing them for success in the 21st century.

Formative Assessment: Transforming Instruction

In 1998, U.K. educators Paul Black and Dylan Wiliam created a stir in the U.S. with their article, "Inside the Black Box." In the article, Black and Wiliam concluded from a synthesis of 250 international studies that formative assessment was a profoundly successful instructional process with an effect size ranging from 0.4 to 0.7—which translates into between 16 and 26 percentile point gains in achievement (Black & Wiliam, 1998b). However, the term "formative assessment" was quickly misappropriated for the frequent use of off-the-shelf tests, with weak ties to curricula (Shepard, 2005). In response to the growing white noise, the Council of Chief State School Officers (CCSSO) convened a formative assessment task force of national experts in 2005 and later formed a state collaborative promoting formative assessment. In 2007, the group disseminated the following definition of formative assessment:

> Formative assessment is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes. (CCSSO, 2012, p. 4)

Chappuis and Stiggins (2002), Wiliam (2007), Wylie (2008), McManus (2008), Heritage (2010), and others have expanded upon the definition of formative assessment, typically citing the following key elements, strategies, or practices in the process:

- teachers ensuring students understand the learning targets and the criteria for success;
- teachers gathering rich evidence of student learning through a variety of means (e.g., observation, questioning, quizzes);
- teachers providing descriptive feedback (instead of grades) on gaps in student learning (which should relate to learning progressions associated with the learning targets);
- teachers and students using the feedback to adjust instruction and learning activities;
- students engaging in self-assessment and meta-cognitive reflection; and
- teachers activating other students as resources.

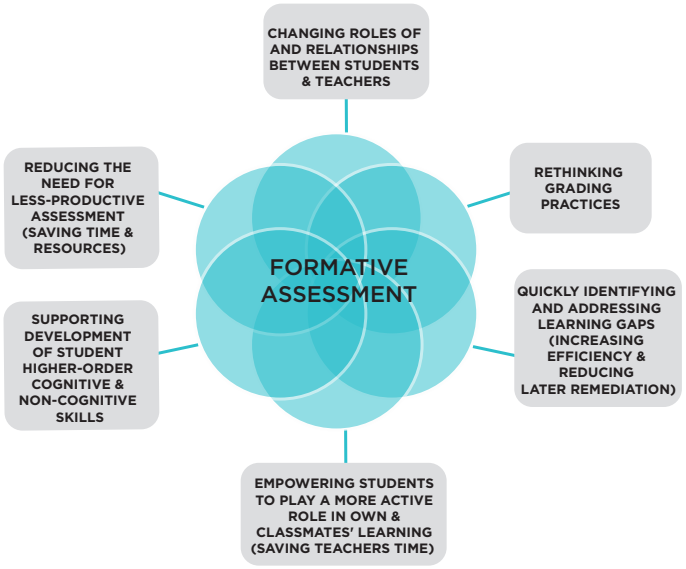
These elements, individually and collectively, have been researched extensively and have been found to help all students grow, with the greatest impact on lower-performing ones (i.e., they help close achievement gaps) (Black & Wiliam, 1998b; Brookhart, 2005; Hattie & Timperley, 2007; and Shute, 2007). In fact, the impact exceeds that of smaller class sizes and matches that of one-on-one tutoring (Black & Wiliam, 1998a).

Formative assessment is both good teaching and good learning, because it empowers students to self-assess and guide their own learning. It can be used for all students, in all grades and subject areas, and for foundational knowledge and higher-order cognitive and non-cognitive skills. Indeed, every school has naturally effective practitioners (who may not apply this name to what they do) in art, music, drama, and career and technical education. How they coach their students to achieve growth and mastery often mimics the essential elements of the formative assessment process.

To reiterate, formative assessment is not simply frequent testing. Nor is it interim or benchmark assessments, such as those provided by publishers or the multi-state assessment consortia. Rather, it is a sequence of instructional steps, one of which involves ongoing monitoring and evidence gathering of student learning related to a particular learning target. This evidence gathering occurs *during instruction* to provide *real-time feedback* to students and teachers to *guide adjustments* they both can make to learning and teaching, and it can be accomplished by a variety of techniques, tools, activities, and measurement instruments besides typical classroom tests and quizzes. Finally, formative assessment is not a silver bullet or a quick-fix solution. It takes time and effort to implement, but when done well, can have, in effect, a multiplier effect on various components of the learning process, as shown in Figure 1 (see p. 7).

These changes alone offer the promise of transforming classrooms and schools, producing equally profound, improved student outcomes (including shrinking achievement gaps), and greatly increasing educational efficiency.

FIGURE 1. MULTIPLIER EFFECT OF FORMATIVE ASSESSMENT



Performance Assessment: Right for the Times

Consider for a moment the process of obtaining a state driver's license. It typically has three components: a multiple-choice test covering basic facts, a period of learning and driving practice, and a behind-the-wheel performance test. All three are important. Knowing the multiple-choice test is coming, a driver candidate digs into the licensing manual, memorizing information on stopping distances, the meanings of signs, and many other rules of the road. But driver and pedestrian safety would be in a sorry state if licenses were granted on the basis of this test alone. Driver candidates need to learn how to put that foundational knowledge to work in the process of actually driving. Achieving mastery takes time, practice, and a lot of formative feedback to prepare driver candidates for the high-stakes summative assessment: demonstrating to an examiner the ability to apply important knowledge and skills to the act of driving itself.

Performance assessment, in effect, applies this same process to classroom learning by requiring students to demonstrate knowledge and skills through some form of product, presentation, or demonstration. It requires students to *apply* knowledge and skills in authentic disciplinary and interdisciplinary tasks related to key aspects of academic learning (Pecheone & Kahl, 2014). Though these tasks may be relatively brief, they nonetheless require students to draw upon and demonstrate deep knowledge and the ability to synthesize, think creatively, and apply what they know in real-world settings. In so doing, performance assessment builds upon foundational knowledge and skills while promoting higher-order thinking and important non-cognitive skills such as collaboration, organization, and initiative. It also provides rich insights into student learning that can inform instructional interventions and provide both summative **and** accountability data. In short, performance assessments enable us to richly, rigorously, and accurately measure what matters most for student growth, promotion, graduation, college, career, and citizenship.

A Brief History of Performance Assessment

Performance assessment is not new. Indeed, it is likely as old as assessment itself. More than 100 years ago, progressive educators encouraged the use of portfolios to measure students' higher-order skills. The authentic-assessment era of the late 1980s and early 1990s saw the adoption of portfolios and performance tasks in several states as part of local—and even statewide—assessment programs that persist today in some school systems. During this period, educators and assessment experts alike learned much about the importance of aligning the assessments with key content and ensuring scoring reliability so results might be used for accountability purposes. Even more recent advances in technology (e.g., electronic portfolios and distributed scoring) can relieve some of the logistical challenges that hampered these efforts in the past, making the use of performance assessment more feasible.

Despite this early progress, we hit a bump on the road to better integrating performance assessments into the learning process when increased annual testing, quick turn-around of test results, and high-stakes accountability prompted states to cut back or eliminate performance-based components of their large-scale assessments, even as the need for deeper learning to prepare students for college, career, and citizenship grew more evident. In recent years, though, the call for college and career readiness and growth of competency-based reforms have prompted more widespread use of performance assessment. As a result, now may be a better time than ever to more significantly embed performance assessment in K–12 education.

The Case for Performance Assessment

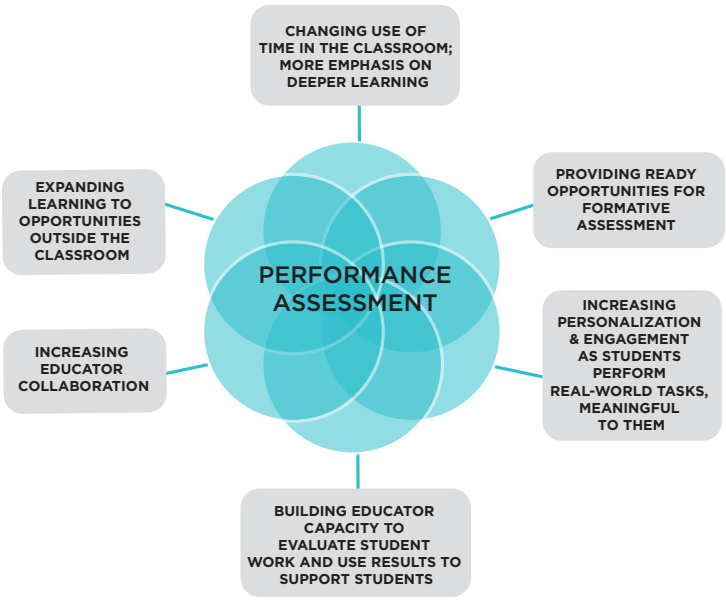
Although the research surrounding performance assessment is not nearly as extensive as that for formative assessment, evidence from multiple sources indicates it focuses instruction on higher-order thinking skills, provides more complete assessment of students' capabilities, increases student engagement, increases teacher buy-in and collaboration, and improves instruction (Darling-Hammond & Wood, 2008). Specifically, Faxon-Mills and colleagues (2013) found that performance-based assessments can drive positive changes in teaching practices, including greater classroom emphasis on critical thinking and real-world problem solving. After the state of Maryland adopted an assessment program with more performance tasks, teachers reported placing greater emphasis on complex problem solving and reasoning in the classroom (Lane, Parke, & Stone, 2002).

A recent synthesis of research by researchers at RAND observed that high-stakes testing can have a positive or a negative effect on teaching and learning, with the former more likely when the assessments are more performance-based and measure higher-order skills and the latter more likely with multiple-choice tests (Soland, Hamilton, & Stecher, 2013). The researchers also commented that performance assessments offer some of the "greatest risks and rewards" when it comes to measuring student learning. On the upside, they can be "most true to life in terms of the demands placed on students," which, in turn, can translate into profound multiplier effects on many aspects of learning, as shown in Figure 2. On the downside, performance assessments are prone to scoring inconsistencies, which has prevented their use for large-scale accountability purposes (Soland, Hamilton, & Stecher, 2013, pp. 37–38). However, as we will describe in this section, by wrapping a structured process around curriculum-embedded performance assessments, it's possible to not only ensure consistency in scoring, but in so doing, improve teachers' instructional practices and professional collaboration while setting, and helping students meet, a high bar for learning.

Putting It All Together: Curriculum-Embedded Performance Assessment

Remember the old television commercials for Reese's peanut butter cups in which two people collide, inadvertently mixing chocolate with peanut butter?

FIGURE 2. MULTIPLIER EFFECT OF PERFORMANCE ASSESSMENTS



When they taste their accidental creation, they find it is better than either ingredient by itself. Similarly, when we integrate formative and performance assessment, we end up with something that is even better than either alone and touches on all three drivers of Daniel Pink's "Motivation 3.0" discussed earlier in this paper. By offering students some *autonomy* in the selection of the performance tasks themselves and the strategies they apply to tackle them, we promote ownership of their learning. By giving real-time feedback during the learning process, we can help students move toward *mastery* of their learning. Finally, by aligning performance tasks with real-world learning, students find more *purpose* to what they are learning and doing.

All three of these components come together in CEPAs (see sample CEPA on p. 9). By embedding performance assessment in curriculum as part of discrete lessons, units, or whole project-based programs, we can promote, measure, and guide deeper student learning. In part, the impact comes from the ability to use formative assessment extensively in the instructional phase, with all of its attendant benefits, and use performance assessment for the deeper engagement and application phases of learning, encouraging and empowering students to carry out richer, more rigorous, and meaningful work. In so doing, we promote students' motivation to learn, an essential ingredient in student success. By relying heavily on the formative assessment process in the instructional phases, CEPAs challenge and support students by giving them a

clear understanding of what success looks like, helping them chart a path to success, encouraging student ownership of the journey, and providing ongoing feedback to motivate students to stretch themselves on assessment tasks. Also, the formative and summative assessment activities within CEPAs are well aligned with one another.

Effective performance tasks are not merely entertaining diversions from schoolwork. In places like New Technology High School in Sacramento or Leadership High School in San Francisco, performance assessment tends to be far more rigorous and demanding of students than anything experienced with traditional selected-response exams. As a student at Leadership High School commented, "At other high schools, it's just 'you passed.' Kids can't tell what they got out of high school. Students here know what they've learned" (Darling-Hammond & Friedlaender, 2008). Moreover, using performance assessment to measure and demonstrate student learning does more than traditional test-based accountability schemes to change teaching practice and improve student learning (Darling-Hammond & Adamson, 2014). Some notable models for using CEPAs range from application for selected standards, as in Ohio's Performance Assessment Pilot Project, to immersive, school-wide programs throughout the year as practiced by several networks, such as schools using Quality Performance Assessment, sponsored by the Boston-based Center for Collaborative Education.

CEPAs are for *all* students. Traditionally, curriculum materials have included enrichment activities that tended to be reserved for use by only the highest achieving students. The CEPA instructional units are designed so that all students benefit from highly engaging, authentic tasks or projects. Furthermore, CEPAs can include collaborative group activities, in which any student can play a role and which provide multiple access points enabling students to be meaningfully involved.

Using CEPA Results for More Meaningful Accountability

As we noted earlier, there is nothing wrong with standardized testing; rather, it's our over-reliance on low-level standarized tests to serve as the sole data point for accountability systems that causes problems. We need a more balanced approach to accountability assessment. We believe that CEPAs offer an effective counterweight to large-scale standardized assessments

Sample CEPA

Below is a brief example of a CEPA. A fully developed CEPA would include content standards and learning targets and offer additional guidance for instruction and assessment, as well as scoring rubrics and sample student work.

Heat Transfer

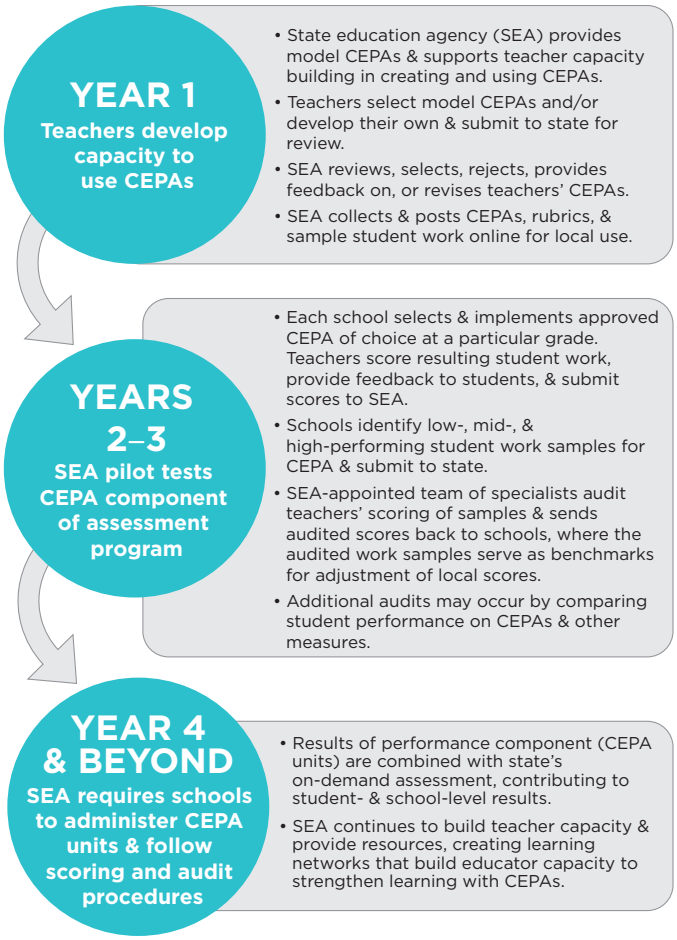
- Activity 1: Students individually or in small groups research methods of heat transfer online. They discuss what they have learned about conduction, convection, and radiation *(student-guided learning).*
- Activity 2: Teachers check student understanding of methods of heat transfer via ungraded quizzes, interviews, or class discussion *(formative assessment evidence gathering, feedback, and adjustment).*
- Activity 3: In small groups, students design and conduct an experiment to determine which of two fabrics better protects against the winter cold. Materials required include tin coffee cans of different sizes (with lids), two different fabrics (e.g., plastic and wool), fasteners, thermometers (thermal probes), timers, and hot water *(performance activity)*.
- Activity 4: Students individually write up a formal lab report of their experiment *(graded summative product).*
- Activity 5: Teachers, via questioning, lead class discussion of how methods of heat transfer played a role in the design and implementation of the research *(formative assessment reflection and reinforcement).*
- Activity 6: Students individually research how a home heating system works and write a paper describing a home heating system and how different methods of heat transfer are involved *(graded summative product).*

by adding (and valuing) a local component in which teachers play an important role. Certainly, we can continue to measure "core" or foundational knowledge with traditional summative tests, but we must also recognize that such tests do not adequately measure applied learning or higher-order skills. CEPAs, however, address both low-level knowledge and deeper learning. Moreover, and perhaps contrary to popular perception, the performance tasks within CEPAs *can* be used for accountability purposes from the classroom to the federal level.

We will not change practice overnight, though. Developing a system of accountability based on CEPAs would likely involve a multi-step, multi-year effort (see Figure 3) involving a great deal of technical work overseen by state education agencies such as developing performance assessment templates and models and vetting and curating CEPAs, as well as training and auditing scorers. It would also require developing significant teacher capacity to deliver CEPAs and evaluate student work. While it may seem daunting, the efforts in Ohio and Boston show that it is possible.

Certainly, this will be hard work and it comes at a time when resources are scarce. Yet it is the *right* work. For starters, in keeping with the truism that what gets tested gets taught, incorporating CEPAs and on-demand performance tasks into state accountability systems would focus more classroom teaching and learning activities on developing students' deeper learning. It could also remove much of the "black box" and mystery that surrounds testing for students and their families by giving them a clearer picture of learning success, along with more timely data on student progress. The accountability system itself would also provide students, parents, taxpayers, higher education institutions, and others with more meaningful measures of student capabilities. Perhaps most importantly, we might turn some of the collateral damage of our current accountability systems into collateral benefits. Instead of compelling teachers to teach to the test, we would build teacher capacity to provide personalized learning experiences with real-world application and deliver more meaningful, real-time feedback to students. At the same time, the process we outlined would provide teachers with opportunities to learn together as professionals from samples of actual student work about how to set—and achieve—high expectations for learning.

FIGURE 3. IMPLEMENTATION OF CEPA COMPONENT IN STATE ACCOUNTABILITY ASSESSMENT



Yes, It Can Be Done

We know these ideas may not be met with immediate enthusiasm. Some may fear that performance assessments would be too time-consuming and impose yet another burden on already stressed teachers. Thus, we advocate for *curriculum-embedded* performance assessments that are not an add-on or distraction but rather the real work of schooling and meant to be included in regular coursework and grades.

Some may argue that teacher-graded performance assessments would be too unreliable. Thus, we propose a system of scoring audits for the high-stakes CEPA component—to ensure that student work products from the performance tasks are scored consistently so the results are reliable and comparable. What teachers learn in this process can be applied to other CEPAs during the year, thus increasing consistency of expectations across all student work.

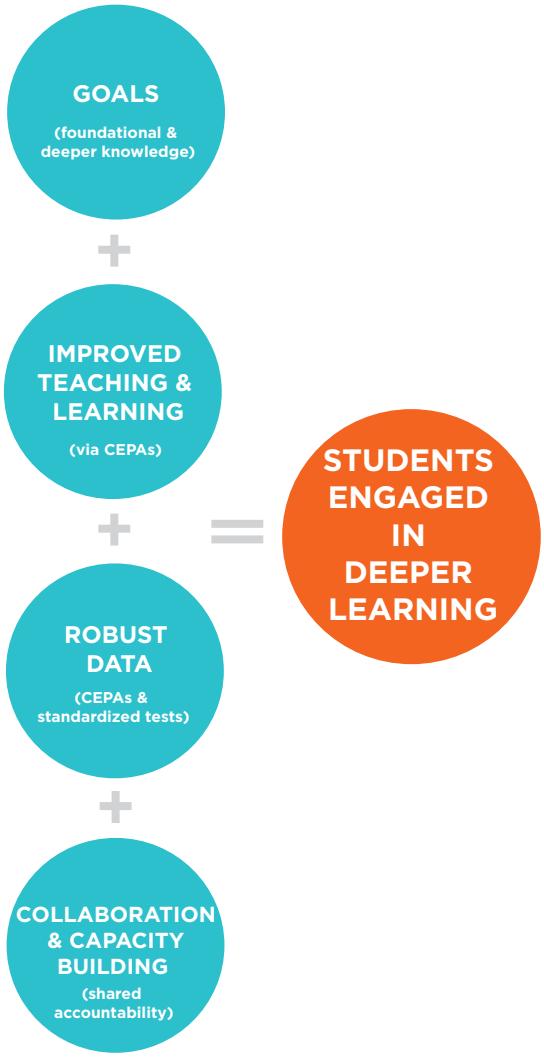
Still others may argue that the entire system would simply be too expensive to administer. However, there are a number of ways we could offset the costs of the 100-plus standardized assessments students currently encounter between kindergarten and high school. Over time, as CEPA use increases, confidence in their reliability grows, and they play a larger role in accountability assessment, the use of many standardized interim assessments we currently administer would be reduced, saving time and money. Moreover, as schools, districts, or states demonstrate quality, growth, and equity, the on-demand accountability testing could be scaled back in terms of testing time because shortened on-demand tests and CEPA components combined would be sufficient for the required technical quality.

Some people have concerns about test security with our current accountability testing systems. To a great extent, security concerns should be less of an issue with CEPAs. Having selected the CEPAs to incorporate in their curricula, school staff will be aware of the content of the CEPAs well before implementing them. However, for summative tasks within CEPAs used for accountability purposes, there would still be directions for teachers to follow, and certifications that the directions were followed, just as there are in traditional accountability testing for security and administrative consistency.

The bottom line is yes, the transition to using CEPA results for accountability purposes can be done. It will take time, effort, and persistence, to be sure. But instead of working harder for diminishing returns, we would be directing our resources and energies to new efforts with the promise of delivering different–and far improved–results for students.

Final Thoughts: A New Formula, Well Worth the Effort

Just as no one argues for scrapping Newtonian physics, we do not advocate eliminating standardized testing or accountability from K–12 education. Rather, as with scientists who have added relativity and quantum mechanics to increase our understanding of the physical world, we advocate a rebalancing of how we use assessment by directing more energy and resources to locally implemented CEPAs to support student growth and deeper learning. In sum, we offer an innovative and more robust formula for measuring learning and driving system success—one that begins with the end in mind and helps students become more engaged learners with



deeper knowledge (see graphic above). We also recognize that external pressure alone is not enough to change the system. As with any adaptive challenge (Heifetz & Laurie, 1997) where the path forward is not clear, we cannot dictate these changes from afar; rather, we must support and encourage teacher collaboration and sharing around actual student work.

As noted earlier, new measures alone are not a quick fix or "silver bullet" solution, but will take time and persistent effort to implement well. Even more importantly, though, they represent a fundamental course correction in education reform—nothing short of a paradigm shift. For too long, we have tried to drive achievement and school accountability with a single-minded focus on a "bottom line" of learning measured

by large-scale, standardized, summative assessments dominated by lower-level items. These are important measures, but they should not be the only measures. By changing this "bottom line" to a more comprehensive picture of student learning as supported and measured by multiple assessment components, including performance assessments, we will begin to change a great deal more, starting with how students and teachers spend their time and perceive their respective roles in the education enterprise.

Ultimately, the promise of CEPAs is that they provide a more motivating, robust, and balanced way to measure student learning. If we believe the maxim that what you test is what gets taught, then these new measures hold the promise of driving many positive changes throughout the system—including better engaging students, supporting deeper learning, encouraging new classroom practices, and supporting greater teacher collaboration. Although better measures alone won't address all of the challenges facing schools, we believe a new formula for measuring student success may be what is most needed to put our nation's schools on a path that breaks through performance ceilings and creates a generation of highly motivated students engaged in deeper learning.

About the Authors

Peter Hofman is an independent consultant who spent 16 years at Measured Progress engaged in mission-driven public policy and outreach initiatives, market research, marketing and communications, strategic planning and partnerships, and special projects.

Bryan Goodwin is President and CEO of McREL International, a non-profit education research and development agency, and author of three books and numerous journal articles, including a regular research column for *Educational Leadership*.

Stuart Kahl is Founding Principal of Measured Progress, Inc., an educational testing company offering products and contracted services for statewide and local assessment programs across the country.

References

Ahn, T., & Vigdor, J. (2013). *Were all those standardized tests for nothing? The lessons of No Child Left Behind.* Washington, DC: American Enterprise Institute.

Ahn, T., & Vigdor, J. (2014, September). *The impact of No Child Left Behind's accountability sanctions on school performance: Regression discontinuity evidence from North Carolina* (Working Paper No. 20511). Cambridge, MA: National Bureau of Economic Research.

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

Black, P., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139–144, 146–148.

Bronson, P., & Merryman, A. (2013). *Top dog: The science of winning and losing.* New York: Twelve.

Brookhart, S. M. (2005, April). *Research on formative classroom assessment*. Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Chappuis, S., & Stiggins, R. J. (2002). Classroom assessment for learning. *Educational Leadership, 60*(1), 40–44.

Council of Chief State School Officers. (2012). *Distinguishing formative assessment from other educational assessment labels*. Washington, DC: Author.

Council of Chief State School Officers, Council of Great City Schools. (2014). Commitments from CCSSO and CGCS on High-Quality Assessments. Retrieved from http://www.ccsso.org/documents/2014/CSSOCGCSAssessmentCommitments10152014.pdf

Csikszentmihalyi, M., Rathunde, K., & Whalen, S. (1993). *Talented teenagers: The roots of success and failure.* New York: Cambridge University Press.

Darling-Hammond, L., & Adamson, F. (Eds.). (2014). *Beyond the bubble test: How performance assessments support 21st century learning*. San Francisco: Jossey-Bass.

Darling-Hammond, L., & Friedlaender, D. (2008). Creating excellent and equitable schools. *Educational Leadership, 65*(8), 14–21.

Darling-Hammond, L., & Wood, G. (2008). *Assessment for the 21st century: Using performance assessments to measure student learning more effectively.* Washington, DC: Forum for Education and Democracy.

Engel, S. (2011). Children's need to know: Curiosity in schools. *Harvard Educational Review, 81*(4), 625–645.

Faxon-Mills, S., Hamilton, L. S., Rudnick, M., & Stecher, B. M. (2013). *New assessments, better instruction? Designing assessment systems to promote instructional improvement*. Santa Monica, CA: RAND.

Gordon Commission on the Future of Assessment in Education. (2013). *A public policy statement*. Princeton, NJ: Educational Testing Service.

Gorski, E. (2014, November 13). Thousands of Colorado high school students refuse to take state tests. *Denver Post*. Retrieved from http://www.denverpost.com/news/ci_26930017/hundreds-colorado-high-school-students-refuse-take-sate

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

Heifetz, R. A., & Laurie, D. L. (1997). The work of leadership. *Harvard Business Review, 75*(1), 124–134.

Heritage, M. (2010) *Formative Assessment and Next-Generation Assessment Systems: Are We Losing an Opportunity?* Paper prepared for the Council of Chief State School Officers, Washington, DC.

Hopkins, D. (2013). *Exploding the myths of school reform.* East Melbourne, Victoria, Australia: Centre for Strategic Reform.

Institute of Education Sciences. (June 2013). *The Nation's Report Card: Trends in Academic Progress* Washington, DC: Author.

Kamenetz, A. (2014). Testing: How much is too much? Retrieved from http://www.npr.org/blogs/ed/2014/11/17/362339421/testing-how-much-is-too-much

Lane, S., Parke, C. S., & Stone, C. A. (2002). The impact of a state performance-based assessment and accountability program on mathematics instruction and student learning: Evidence from survey data and school performance. *Educational Assessment, 8*(4), 279–315.

Lazarín, M. (2014). *Testing overload in America's schools.* Washington, DC: Center for American Progress.

Levitt, S. D., List, J. A., Neckerman, S., & Sadoff, S. (2012). *The behavioralist goes to school: Leveraging behavioral economics to improve educational performance* (NBER Working Paper Series No. 18165). Cambridge, MA: National Bureau of Economic Research.

McManus, S. (2008). *Attributes of Effective Formative Assessment.* Paper prepared for the Formative Assessment for Teachers and Students (FAST) State Collaborative in Assessment and Student Standards (SCASS) of the Council of Chief State School Officers (CCSSO), Washington, DC.

MetLife, Inc. (2012). *The MetLife survey of the American teacher: Challenges for school leadership.* Retrieved from https://www.metlife.com/assets/cao/foundation/MetLife-Teacher-Survey-2012.pdf

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

Nichols, S. L., Glass, G. V., & Berliner, D. C. (2012). High-stakes testing and student achievement: Updated analyses with NAEP data. *Education Policy Analysis Archives, 20*(20).

Organisation for Economic Cooperation and Development (OECD). (2012). Programme for International Student Assessment (PISA): Results from PISA 2012: United States. Retrieved from http://www.oecd.org/pisa/keyfindings/PISA-2012-results-US.pdf.

Pecheone, R., & Kahl, S. (2014). Where are we now: Lessons learned and emerging directions. In L. Darling-Hammond & F. Adamson (Eds.), *Beyond the Bubble Test: How Performance Assessments Support 21st Century Learning* (pp. 53–91). San Francisco, CA: Jossey-Bass.

Pelligrino, J. W., & Hilton, M. (Eds.) (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century.* Washington, DC: National Academies Press.

Pianta, R. C., Belsky, J., Houts, R., & Morrison, F. (2007). Opportunities to Learn in America's Elementary Classrooms. *Science, 315*(5820), 1795–1796.

Pink, D. H. (2009). *Drive: The surprising truth about what motivates us*. New York: Riverhead Books.

Shepard, L. A. (2005). *Will Commercialization Enable or Destroy Formative Assessment?* Paper presented at the ETS Invitational Conference 2005: The Future of Assessment, Shaping Teaching and Learning, New York.

Shute, V. J. (2007). *Focus on formative assessment* (ETS Report No. RR-07-11). Princeton, NJ: Educational Testing Service.

Soland, J., Hamilton, L., & Stecher, B. (2013). *Measuring 21st century competencies – guidance for educators.* Retrieved from http://asiasociety.org/files/gcen-measuring21cskills.pdf

Wiliam, D. (2007). Keeping learning on track: Classroom assessment and the regulation of learning. In F.K. Lester (Ed.), *Second Handbook of Mathematics Teaching and Learning.* Greenwich, CT: Information Age Publishing.

Wylie, E. C. (2008). *Formative Assessment: Examples of Practice.* Paper prepared for the Formative Assessment for Teachers and Students (FAST) State Collaborative in Assessment and Student Standards (SCASS) of the Council of Chief State School Officers (CCSSO), Washington, DC.