Research Says / Better Tests Don't Guarantee Better Instruction

By Bryan Goodwin

Educational Leadership Vol 71, Number 6
March 2014

With the approach of the new assessments associated with the Common Core standards, some educators may hope that a new day is dawning. These educators may reason that by providing more thoughtfully constructed measures of higher-order thinking, the new assessments could lead to better classroom instruction.

After all, we know that "what you measure is what you get." If we only test basic skills, that's what teachers will teach and what students will learn. This lesson was supported by a recent review of research on high-stakes tests. Faxon-Mills, Hamilton, Rudnick, and Stecher (2013) found that these tests, which typically measured basic knowledge, drove teachers to spend more effort "promoting basic skills while devoting less attention to helping students develop creativity and imagination" (p. 16). It is reasonable to expect, then, that better tests—those that require students to master more difficult content and demonstrate critical thinking—will drive better learning outcomes.

Our cycle of pinning our hopes on the power of student outcome measurement calls to mind the 1993 comedy Groundhog Day, in which Bill Murray plays a self-centered TV weatherman who finds himself snowbound in Punxsutawney, Pennsylvania, doomed to repeat the same day until he finally changes his ways. In light of the uneven track record of previous test-driven reforms, wary educators might reasonably ask, Will better high-stakes assessments really change anything? Or will we just experience Groundhog Day all over again?

Confronting Campbell's Law

Unfortunately, any attempt to drive education improvement with high-stakes testing and accountability may have a fundamental flaw. In the 1970s, Donald Campbell, then president of the American Psychological Association, theorized that "the more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor" (1976, p. 49). In other words, according to Campbell's law, the higher the stakes attached to any measure, the less valid that measure becomes.

Consider the dramatic gains reported over the years on some states' accountability assessments. If these gains reflect true increases in student learning, we would expect to see these states' students demonstrating similar gains on other measures. The troubling reality, though, is that gains on high-stakes tests do not appear to translate into gains on lower-stakes assessments.

Harvard researcher Brian Jacob (2002), for example, conducted an in-depth analysis of test scores in Chicago Public Schools during a period (1993–2000) when student achievement increased by .30 standard deviations (12 percentile points) in mathematics and .20 standard deviations (8 percentile points) in reading. During that same period, however, Jacob found that the achievement of the same students on a low-stakes statewide exam that ostensibly measured the same learning standards dropped significantly. This finding prompted him to conclude that "despite its increasing popularity within education, there is little empirical evidence on test-based accountability" (p. 4).
Similarly, a more recent analysis of a decade of trend data in 25 states (Nichols, Glass, & Berliner, 2012) found that high-stakes testing pressure on statewide assessments under No Child Left Behind (NCLB) did not translate into better student achievement on the low-stakes National Assessment of Educational Progress (NAEP). Only in 4th grade math did NAEP scores appear to improve after NCLB; but they did so at a slower rate than they had before NCLB. Moreover, in states with the highest stakes attached to test results, NAEP reading scores for students living in poverty appeared to decline.

In other words, although high stakes may cause test scores to rise on a particular assessment, those scores may not reflect true gains in student learning. Rather, the gains may reflect something else—arguably, how well teachers boosted students' test-taking abilities or narrowed instruction to the knowledge captured on the test.

**Will This Time Be Different?**

What happens, though, if we administer better tests—for example, those that employ performance tasks to measure critical thinking and student application of knowledge? Do teacher practices and student learning change in positive ways?

Here, there appears to be some cause for optimism. Faxon-Mills and colleagues (2013) found that performance-based assessments—like those promised in the new Common Core assessment systems—do have the potential to drive positive changes in teaching practices, including encouraging greater classroom emphasis on critical thinking and real-world problem solving.

However, by themselves, performance assessments do not guarantee better teaching or learning. After Maryland adopted an assessment that included more performance tasks, teachers reported placing greater emphasis on complex problem solving and reasoning in the classroom (Lane, Parke, & Stone, 2002)—but a follow-up survey of students (Parke & Lane, 2007) found that most were still experiencing a heavy dose of basic short-answer recall questions and textbook-based teaching.

**A Cure for Campbell's Law**

According to Campbell (1976), one way to prevent an indicator from becoming corrupted or distorted is to employ multiple measures of performance. Another way, research suggests, is to emphasize formative data—low-stakes classroom assessments created by teachers to guide instruction, which can have a strong, positive influence on student performance and motivation (Wiliam & Thompson, 2007).

The key lies in leveraging the daily or weekly nature of these assessments to guide real-time changes to classroom instruction (Wiliam & Thompson, 2007). Therefore, formative assessments should not be confused with so-called interim or benchmark assessments, which are often just large-scale assessments repackaged as monthly (or longer cycle) tests. Research shows that such tests do little or nothing to improve instruction (Popham, 2006).

**Avoiding Groundhog Day**

Large-scale assessment is not necessarily a bad idea; it can provide useful comparative data to identify bright spots and guide system change. It's worth noting that in the same paper in which Donald
Campbell (1976) advanced Campbell’s law, he also strongly advocated using evaluation and data to guide system change.

So designing large-scale tests that measure higher-order student outcomes is certainly a worthwhile first step. However, research suggests that better tests, although necessary, are probably not sufficient to effect change. Unless any new assessment system is accompanied by systems of multiple indicators to mute the distortions implied by Campbell's law, teacher capacity building to support better classroom practices, and greater emphasis on short-cycle formative classroom assessment to guide instruction, we may be doomed to repeat, in *Groundhog Day*–fashion, the frustrations of the past.

**References**


*Bryan Goodwin* is chief operating officer at McREL, Denver, Colorado. He is the author of *The 12 Touchstones of Good Teaching: A Checklist for Staying Focused Every Day* (ASCD, 2013).