Teacher Evaluation as Trojan Horse: The Case for Teacher-Developed Assessments

Derek C. Briggs

University of Colorado

March 1, 2013

In his focus article "How is Testing Supposed to Improve Schooling?" Ed Haertel distinguishes between seven uses of educational tests as a function of the intended action and what or who will be influenced by the intended action. He then applies Mike Kane's interpretive argument approach (Kane, 2006) as a basis for speculating about the validity of intended actions. A recurring theme in the article is that validating the premise that testing has improved schooling will require much more attention to issues of score extrapolations and decision-making. In this comment I expand upon the section of Haertel's paper that focused on testing for accountability. More specifically, I take up the management role of testing within the context of recent policy shifts towards high-stakes teacher evaluation.

In teacher evaluation, the chain of reasoning that connects testing to improvements in schooling goes something like this:

1. No school-specific variable has a bigger potential impact on student learning than teachers.
2. If teachers are being evaluated, then the evaluation criteria should include both evidence regarding the quality of a teacher's classroom practices, and evidence regarding the amount that students have learned.

3. The best way to infer what students have learned is to test them on multiple occasions.
4. Teachers who are consistently associated with students who perform worse than some normative or criterion-referenced expectation should be given additional professional development, or fired.
5. In the long run, schooling improves because (a) ineffective teachers can be identified and counseled out the system, (b) effective teachers can be rewarded on the basis of merit rather than seniority, and as a consequence of (a) and (b), teacher are motivated to work harder.

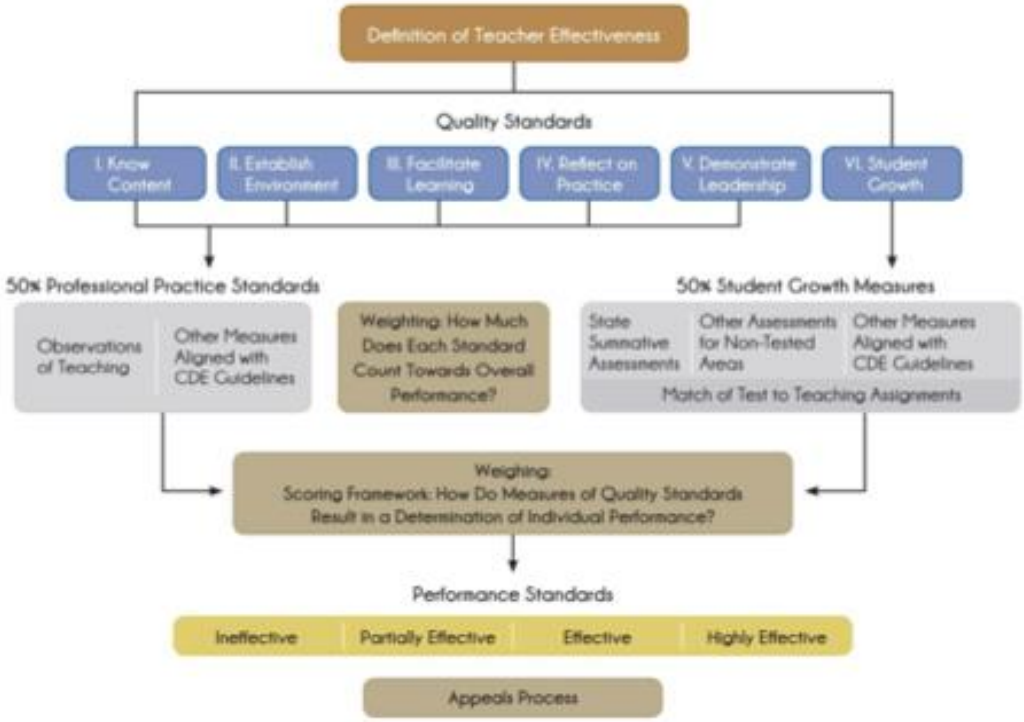
I would call 1-5 above the economic model of teacher evaluation (c.f., Gordon, Kane & Staiger, 2006). So in the example above, *student* testing is a means to an end in that it makes it possible to evaluate teachers with respect to the value-added they appear to contribute to students. However, in his article Haertel makes the testing of students and evaluation of teachers synonymous by conceptualizing value-added model estimates themselves as the "scoring" of teachers.

One of Haertel's concerns is that not enough attention has been paid in critiquing the economic model to the questions of how strongly value-added scores are related to broader notions of teaching quality. I would disagree with this position--though I suppose it depends on what Haertel means by "enough." The $52-million Gates-funded Measures of Effective Teaching project devoted the past 3 years to exploring this very issue. The ongoing Understanding Teaching Quality project (featuring researchers from ETS, RAND and the University of Michigan) continues to explore this issue. And as local teacher evaluation systems go to scale, there surely will be countless other examples

in which the relationships among multiple measures of teaching quality are being examined empirically. What we have learned from this work so far is that value-added scores tend to be weakly correlated with direct observations of teaching practice (MET paper). Interestingly, value-added scores are only moderately correlated across tests in different subject domains (e.g., math-reading), and even across tests from the same subject domain (e.g., math computation-math reasoning) (Kane & Staiger, 2010; Lockwood et al., 2007; Papay, 2011). It isn't entirely clear how this evidence should be interpreted with respect to a validity argument. Different measures of teaching quality are related, but not very strongly.

One way this could be taken is as an argument for the importance of incorporating multiple measures into the evaluation of teachers. An example of this can be seen in Figure 1, which contains the model teacher evaluation framework that was proposed by Colorado's State Council on Educator Effectiveness. Colorado's law mandating the annual evaluation of teacher (and principals) will be go into full implementation statewide as of the 2014-15 school year. All school districts will be required to classify public school teachers into one of four categories: ineffective, partial effective, effective and highly effective.

Figure 1. Colorado's Model Teacher Evaluation Framework



These ratings will be based on two distinct sources of evidence: 50% evidence of effectiveness with respect to professional practice standards, and 50% evidence with respect to growth in student learning outcomes. Within these two sources there will be multiple measures based on direct observations of teachers, and multiple measures based on test scores and student growth objectives.

There are at least two reasons to be pessimistic that the economic model will improve schooling purely through what Haertel describes as "educational management." Both reasons anticipate the potential negative consequences of using testing for high-stakes decision-making. The first is that teachers will teach to the test and this will lead to a distortion of the curriculum or even outright cheating. The second is that the tests are not the same for all teachers. In particular, although the observation protocols used to rate teachers on professional practice standards are common, only about one-third of K-

12 teachers have students that take the state's large-scale assessments in math, English Language Arts and science. This could lead to the perception that evaluations are unfair because some measures are more valid than others, and cause a decrease in teacher morale and motivation. In the remainder of this comment I will focus on two prospects for mitigating these negative consequences.

Building Better Externally Mandated Tests

Let's imagine the ideal scenario for a hypothetical state. In 2010 the state adopted the Common Core of State Standards (CCSS) as the K-12 targets for learning in mathematics and English Language Arts. Since the advent of test-based accountability, it could be reasonably argued that the CCSS represent the most thoughtful attempt to develop standards that are coherent, specific and meaningful. No one who has read them could argue that they represent another example of standards that are an "inch deep and a mile wide." Having adopted the standards, staff from the state's department of education begin working with school districts to design curriculum and instruction that will be aligned with these standards. While this cannot happen overnight, the state puts in place a four-year plan to phase in new curricular guidelines and accompanying professional development. Starting with the 2014-15 school year, the state will be adopting a testing system developed by the Smarter-Balanced Assessment Consortium (SBAC.

The expressed purpose of SBAC is to develop assessments that are "valid, support and inform instruction, provide accurate information about what students know and can do, and measure student achievement against standards designed to ensure that all

students gain the knowledge and skills needed to succeed in college and the workplace." The SBAC system will have three components:

1. Computer adaptive summative assessment that will be administered during the last 12 weeks of the school year (mandatory)
2. Interim assessments that can be used to predict student performance on the summative assessment while also providing feedback on student progress (optional)
3. Formative assessment resources to help teachers diagnose and respond to the needs of their students as they teach the content of the CCSS (optional)

Even though these last two components are optional, our hypothetical state has signed on to them because its educational stakeholders believe strongly in striking a balance between assessment for formative and summative purposes. The summative component of the assessment is a dramatic change from the previous assessment the state had been using since NCLB. Not only is it targeted to the CCSS, but it places an emphasis on college and career readiness rather than minimum proficiency. To pull this off, the SBAC tests will include performance-based components in both math and ELA that emphasize the sorts of cognitive skills that are hard to elicit using traditional multiple-choice items.

In other words, if teachers teach to the test, it means they would need to, among other things, teach students to integrate knowledge and skills across multiple content standards, and demonstrate the ability to critically analyze and synthesize information presented in a variety of formats. If SBAC were able to deliver on its promise to design a testing system premised on eliciting these sorts of practices, and if the test items could be

written to reflect real-world tasks while also represent content that is relevant and meaningful to students, than teaching to the test would be a desirable outcome.

Although this is a tall order, I am young and optimistic (some might say naïve) enough to believe that even though the consortia-developed assessments are unlikely to meet all the grand ambitions that motivated them, they are likely to represent a distinct improvement over the large-scale assessments that preceded them.

Giving Teachers Ownership of Locally Developed Tests

Although building tests worth teaching to is important, even in a best case scenario this would only apply to one-third of teachers. Even if a magic wand could be waved and state consortia could develop high quality tests for all grades and subjects from Kindergarten through the 12th grade, high-stakes teacher evaluation alone would still be unlikely to lead to significant improvements in schooling via the economic model. The reason for this is simple. Teaching and learning is, at heart, a local and personal experience. Once the door to a classroom is closed, it's just the teacher and the students, and unless a teacher has chosen to internalize the premise behind a system of educational accountability (with its emphasis on growth in student learning as a guiding beacon), that system will not promote the kind of reflection and dialog that would be necessary to improve the practice of teaching. In short, if teachers and principals do not believe the evaluation process is fair and authentic, if they only perceive it is as external red tape to which they must comply, there will be no buy-in, and the reform will crumble. It will

crumble just like the many test-based reforms intended to improve schooling that have come and gone in the past (Shepard, 2008).

I think the only way to overcome this hurdle of buy-in is to give all teachers—even those for whom the consortia-developed summative tests apply—ownership over one or more of the test outcomes that will be used to make inferences about student growth. It is in this sense that the fact that a majority of teachers teach in "untested" subjects could be viewed as a blessing rather than a curse, an opportunity rather than an obstacle. The idea I am promoting here harkens back to the concept of an embedded assessment system, the principles for which were clearly framed in a paper by Wilson & Sloane (2001). Wilson & Sloane argue that four elements must be in place before teachers can be expected to use assessments for formative purposes

1. Involved in the process of collecting and selecting student work.
2. Able to score and use the results immediately—not wait for scores to be returned several months later.
3. Able to interpret the results in instructional terms.
4. Able to have a creative role in the way that the assessment system is realized in their classrooms.

The Denver Public School system in Colorado has for some time now implemented an approach for using assessments to gauge student growth that has the *potential* to fulfill these elements. The approach is known as setting student growth objectives (SGOs, also sometimes described as student learning objectives, SLOs). The basic idea is that for each student in a teacher's class, a teacher should administer a pre-test at the outset of the school year to figure out what it is that students appear to know

and be able to do with respect to key concepts in the curriculum. Using this information, a teacher should set aspirational learning targets for each student, and then enact a plan to get them there. At the end of the year, the teacher should test the students again, compare aspirational growth to actual growth, and then reflect upon on the process so that it can be improved upon in the following year.

Decoupled from its role as evidence in a high-stakes teacher evaluation system, I suspect no one would object to the desirability of teachers using SGOs in the manner described above. The problem to be overcome is that if the SGOs are based on tests that have been externally mandated, teachers are more likely to view the process as a compliance-based hoop they need to jump through than an opportunity to think more deeply about what they are teaching and what kids are actually learning. In practice, the execution of setting SGOs has lagged far behind the noble theory of action outlined above, but not, I would speculate, because teachers are cynical. Rather, it is because the use of tests to make inferences about growth rather than status represents a cultural shift that is not just new to teachers, but also to psychometricians and commercial test developers alike.

There are some templates out there that school districts can learn from, including the work by Bob Mislevy and colleagues on evidence-centered design, and Mark Wilson and colleagues on the BEAR Assessment System. In addition, the innovations in the burgeoning research literature on learning progressions in science and learning trajectories in math could provide a model for how to build assessments for the explicit purpose of evaluating growth. Sketching out in more detail how it might be possible to

achieve formative ends within a summative system will have to be a matter for another paper.

Involving all teachers in choosing the tests by which they and their grade-level colleagues will be evaluated strikes me as the key leverage point for empowering teachers and drawing on their professional knowledge. In a best case scenario, teacher evaluation might be the Trojan horse that smuggles in the principles of embedded assessment systems into active classroom use. If this were to happen, then it might be possible for testing associated with teacher evaluation not only to improve schooling through better educational management, but through genuine improvements to the teaching and learning that goes on behind classroom doors.

References

Gordon, R., Kane, T., & Staiger, D. (2006). Identifying effective teachers using performance on the job. Discussion Paper 2006-01. The Brookings Institution.

Kane, M. (2006). Validation. In R. Brennan (Ed.) *Educational Measurement*, 4th Edition (pp. 17-64). Westport, CT: Praeger.

Kane, T., & Staiger, D. (2010). Learning about teaching. Research report for the Measures of Effective Teaching project.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V. N., & Martinez, J. F. (2007). The Sensitivity of Value-Added Teacher Effect Estimates to Different Mathematics Achievement Measures. Journal of Educational Measurement, 44(1), 47–67.

Papay, J. P. (2011). Different tests, different answers: the stability of teacher value-added estimates across outcome measures. American Educational Research Journal, 48(1), 163–193.

Shepard, L. A. (2008). A brief history of accountability testing, 1965-2007. In K. Ryan & L. Shepard (eds) *The Future of Test-Based Educational Accountability*. Routledge: New York, NY.

Wilson, M. & Sloane, K. (2001). From principles to practice: an embedded assessment system. Applied Measurement in Education, 13(2), 181-208.