Passing Muster Fails Muster?

(An Evaluation of Evaluating Evaluation Systems)

Bruce Baker
Associate Professor
Department of Educational Theory, Policy and Administration
Rutgers, The State University of New Jersey
Bruce.baker@gse.rutgers.edu

The Brookings Institution has now released its web based version of *Passing Muster,* including a nifty calculation tool for rating teacher evaluation systems. Unfortunately, in my view, this rating system fails muster.

The awkward issue here is that this brief and calculator are prepared by a truly exceptional group of scholars, and not just reform-minded pundits. It strikes me that we technocrats have started to fall for our own contorted logic – that the available metric is the true measure – and the quality of all else can only be evaluated against that measure. We've become myopic in our analysis, and we've forgotten all of the technical caveats of our own work, simply assuming the technical caveats of any/all alternatives to be far greater.

Beyond all of that, I fear that technicians working within the political arena are deferring judgment on important technical concerns that have real ethical implications. When a technician knows that one choice is better (or worse) than another, one measure or model better than another, and that these technical choices affect real lives, the technician should – MUST – be up front/honest about these preferences.

To that effect, I have two major concerns about the rating system offered in *Passing Muster*:

First, the authors explain their (lack of) preferences for specific types of evaluation systems as follows:

"Our proposal for a system to identify highly-effective teachers is agnostic about the relative weight of test-based measures vs. other components in a teacher evaluation system. It requires only that the system include a spread of verifiable and comparable teacher evaluations, be sufficiently reliable and valid to identify persistently superior teachers, and incorporate student achievement on standardized assessments as at least some portion of the evaluation system for teachers in those grades and subjects in which all students are tested."

That is, a district's evaluation system can consider student test scores to whatever extent they want, in balance with other approaches to teacher evaluation. The logic here is a bit contorted

from the start. The authors explain what they believe are necessary components of the system, but then claim to be agnostic on how those components are weighted.

But, if you're not agnostic on the components, then saying you're agnostic on the weights is not particularly soothing.

Clearly, they are not agnostic on the components or their weights, because the system goes on to **evaluate the validity of each and every component based on the extent to which that component correlates with the subsequent year value-added measure.** This is rather like saying, "We remain agnostic on whether you focus on reading or math this year, but we are going to evaluate your effectiveness by testing you on math." Or more precisely, "We remain agnostic on whether you emphasize conceptual understanding and creative thinking this year, but we are going to evaluate your effectiveness on a pencil and paper, bubble test of specific mathematics competencies and vocabulary and grammar."

Second, while hanging ratings of evaluation systems entirely on their correlation with "next year's value added," the authors choose to again remain agnostic on the specifics for estimating the value-added effectiveness measures. That is, as I've blogged in the past, the authors express a strong preference that the value added measures be highly correlated from year to year, but remain agnostic as to whether those measures are actually valid, or instead are highly correlated mainly because the measures contain significant consistent bias – bias which disadvantages specific teachers in specific schools – and does so year after year after year!

Here are the steps for evaluating a teacher evaluation system as laid out in Passing Muster:

> Step 1: Target Percentile of True Value Added
> Step 2: Constant factor (tolerance)
> Step 3: Correlation of teacher level total evaluation score in current year, with next year value added
> Step 4: Correlation of non-value added components with next year's value added
> Step 5: Correlation of this year's value added with next year's value added
> Step 6: Number of teachers subject to the same evaluation system used to calculate correlation in step 3 (a correlation with next year's value added!)
> Step 7: Number of current teachers subject to only the non-value added system

In formal terms, their system is all reliability and no validity (or, at least, inferring the latter from the former).

But, rather than simply having each district evaluate its own evaluation system by correlating its current year ratings with next year's value-added, the Brookings report suggests that states should evaluate district teacher evaluation systems by measuring the extent that district teacher evaluations correlate with a state standardized value-added metric for the following year.

But again, the authors remain agnostic on how that model should/might be estimated, favoring that the state level model be "consistent" year to year, rather than accurate. After all, how could districts consistently measure the quality of their evaluation systems if the state external benchmark against which they are evaluated is not consistent?

As a result, where a state chooses to adopt a consistently biased statewide standardized value-added model, and use that model to evaluate district teacher evaluation systems, the state in effect backs districts into adopting consistently biased year-to-year teacher evaluations… that have the same consistent biases as the state model.

The report does suggest that in the future, there might be other appropriate external benchmarks, but that:

> Currently value-added measures are, in most states, the only one of these measures that is available across districts and standardized. As discussed above, value-added scores based on state administered end-of-year or end-of-course assessments are not perfect measures of teaching effectiveness, but they do have some face validity and are widely available.

That is, value-added measures – however well or poorly estimated – should be the benchmark for whether a teacher evaluation system is a good one, simply because they are available and we think, in some cases, that they may provide meaningful information (though even that remains disputable -- to quote Jesse Rothstein's review of the Gates/Kane Measures of Effective Teaching study: "In particular, the correlations between value-added scores on state and alternative assessments are so small that they cast serious doubt on the entire value-added enterprise." See: http://nepc.colorado.edu/files/TTR-MET-Rothstein.pdf).

I might find some humor in all of this strange logic and circular reasoning if the policy implications weren't so serious.