Gates' Measures of Effective Teaching Study: More Value-Added Madness

By Justin Baeder on December 21, 2010 3:16 AM | 7 Comments | Recommend

@eduleadership 2

The Measures of Effective Teaching project, funded to the tune of $45 million by the Gates Foundation, has **released its first of four reports**. While the report is full of intelligent insights, it makes a number of astounding logical leaps to justify the use of value-added teacher ratings, and I can already see how the study will be used to make the case for sloppy valued-added teacher evaluation systems.

This is an ambitious study, and a very well-designed one at that. I can't imagine a better-funded or better-designed study of teacher effectiveness measures; top-notch researchers and the use of five different measures will doubtless make this one of the stronger studies of its type. However, one of its fundamental premises is deeply flawed, and this affects the conclusions drawn throughout the report:

> Second, any additional components of the evaluation (e.g., classroom observations, student feedback) should be demonstrably related to student achievement gains. p. 5

This assumption would make sense if "student achievement gains" were as legitimate and stable a construct as the study asserts. Certainly it makes sense to evaluate teachers based on how well they improve student learning, but the assumption that we can isolate teacher effects from all of the other influences on student test performance has not been borne out by research to date on value-added measurement (VAM), including the research done in the MET study itself.

In fact, every indication is that it will continue to remain impossible to isolate teacher effects from such other influences as:

- Nonrandom assignment of students to teachers, which can happen due to purposeful assignment of struggling students to stronger teachers, or due to unintentional scheduling oddities (for example, if Advanced Band is offered when I teach Algebra II, I'm going to get non-band students in my Algebra II class, who may be different in significant ways from band students)
- Other teachers who serve the same students, e.g. special education, gifted & talented, academic support, or other subject-area teachers whose effects may spill over into other classes
- Class composition, which every teacher will tell you varies from year to year despite attempts to "balance" classes. While it's the teacher's job to create an environment conducive to learning, students do play an important role in determining the culture of a class, which can have a significant impact on learning

Despite attempts to control for such extraneous variables in value-added measurement, there is strong empirical evidence that "student achievement gains" are not stable from year to year nor as the MET report notes, even between different sections of the same subject taught at the same time by the same teacher. As **this EPI briefing paper notes**,

> there is broad agreement among statisticians, psychometricians, and economists that student test scores alone are not sufficiently reliable and valid indicators of teacher effectiveness to be used in high-stakes personnel decisions, even when the most sophisticated statistical applications such as value-added modeling are employed.

Specifically,

> One study found that across five large urban districts, among teachers who were ranked in the top 20% of effectiveness in the first year, fewer than a third were in that top group the next year, and another third moved all the way down to the bottom 40%. Another found that teachers' effectiveness ratings in one year could only predict from 4% to 16% of the variation in such ratings in the following year. Thus, a teacher who appears to be very ineffective in one year might have a dramatically different result the following year. The same dramatic fluctuations were found for teachers ranked at the bottom in the first year of analysis. This runs counter to most people's notions that the true quality of a teacher is likely to change very little over time and raises questions about whether what is measured is largely a "teacher effect" or the effect of a wide variety of other factors.

These cited studies are not the final word, of course, so it would be reasonable for the MET study to try to improve value-added measurements in order to try to obtain more stable year-to-year ratings.

However, that does not appear to be part of the design; instead, first-year value-added scores will be treated as the "truth" against which all other measures of teacher effectiveness will be judged. But how good is this "truth"?

Let's see what the MET study found about the stability of its own value-added measures. Two types of data were available: comparisons between different sections of the same subject taught by the same teacher in the same year, and comparisons between the same class taught by the same teacher in two different years. The report states that

> When the between-section or between-year correlation in teacher value-added is below .5, the implication is that more than half of the observed variation is due to transitory effects rather than stable differences between

> teachers. That is the case for all of the measures of value-added we calculated. We observed the highest correlations in teacher value-added on the state math tests, with a between-section correlation of .38 and a between-year correlation of .40. The correlation in value-added on the open-ended version of the Stanford 9 was comparable, .35. However, the correlation in teacher value-added on the state ELA test was considerably lower--.18 between sections and .20 between years.

In other words, the value-added score for one math class only has a 40% correlation with the same teacher's score for another class, whether taught at the same time or in a different year. For ELA, the correlation is only about 20%.

You'd think that the researchers would at this point give up on value-added and start looking for more reliable measures. Instead, we're treated to a full paragraph of logical gymnastics and implication-avoidance (emphasis mine):

> Does this mean that there are no persistent differences between teachers? Not at all. *The correlations merely report the proportion of the variance that is due to persistent differences between teachers*. Given that the total (unadjusted) variance in teacher value-added is quite large, the implied variance associated with persistent differences between teachers also turns out to be large, despite the low between-year and between-section correlations. For instance, the implied variance in the stable component of teacher value-added on the state math test is .020 using the between-section data and .016 using the between-year data. Recall that the value- added measures are all reported in terms of standard deviations in student achievement at the student level. Assuming that the distribution of teacher effects is "bell-shaped" (that is, a normal distribution), this means that *if* one could accurately identify the subset of teachers with value-added in the top quartile, they would raise achievement for the average student in their class by .18 standard deviations relative to those assigned to the median teacher. Similarly, the worst quarter of teachers would lower achievement by .18 standard deviations. So the difference in average student achievement between having a top or bottom quartile teacher would be .36 standard deviations. That is far more than one-third of the black-white achievement gap in 4th and 8th grade as measured by the National Assessment of Educational Progress--closed in a single year!

So we've gone from "these results are highly unstable" to "we can eliminate the achievement gap in one year!" in the space of a single paragraph. Please leave a comment if I'm misinterpreting this part of the study, but it seems to me that if your measure is only a 20% predictor of *itself*, you don't have a meaningful measure at all. It's certainly true that some teachers are much, much better than others, but forgive me if I'm hesitant to trust a measure that is potentially wrong 80% of the time.

Earlier, the authors describe their plan for handling this risk: simply give VAM less weight by

> scaling down (or up) the value-added measures themselves. But that's largely a matter of determining how much weight should be attached to value-added as one of multiple measures of teacher effectiveness. p. 4

Let me get this straight: If I choose to evaluate you on the basis of a coin toss, which is totally random, I know I'll be wrong 50% of the time. Therefore, the coin toss is a valid evaluation tool provided that it only counts for 50% of your overall evaluation.

Please tell me I'm reading this wrong my background in statistics barely qualifies as graduate-level, and I'm certainly not a VAM expert. But I think I'm interpreting the authors' argument correctly.

Interestingly, student ratings are much more stable between classes and from year to year their correlation is on the order of 67%. If anything, this first MET report provides good evidence that simply asking students about their teachers is a much better idea than going through both statistical and logical gymnastics to obtain a VAM score.

In its closing section, the report argues that VAM is useful even though its predictive power is incredibly weak. Now, if you wanted to report the utter unreliability of VAM as good news, how would you do it? The authors take this tack:

> Two types of evidence student achievement gains and student feedback do seem to point in the same direction, with teachers performing better on one measure tending to perform better on the other measures. p. 31

In other words, the correlations between VAM and other forms of assessment are ridiculously weak, but hey, at least they're not negative. The report goes on to say that

> many people simply forget all the bad decisions being made now, when there is essentially no evidence base available by which to judge performance. Every day, effective teachers are being treated as if they were the same as ineffective teachers and ineffective teachers are automatically granted tenure after two or three years on the job. Given that we know there are large differences in teacher effects on children, we are effectively mis-categorizing everyone when we treat everyone the same. Value-added data adds information. Better information will lead to fewer mistakes, not more. Better information will also allow schools to make decisions which will lead to higher student achievement.

Wow. Notice the careful wording: "Value-added data adds information. Better information..." Does it say that value-added data *is* better information? No, because clearly it's not. If and when we're able to obtain better information (e.g. from student ratings or more rigorous observation methods), we should certainly incorporate it into teacher evaluations. But for now, VAM doesn't give us anything useful just a contextless number accompanied by a false sense of certainty.

In the end, what does this report tell us? Let's look at the problem the study is intended to solve: More than 99% of teachers are rated "satisfactory" every year, which should seem incredible to even the most ardent supporter of teachers. There is no doubt that principals, myself included, need to do a far better job of identifying underperforming teachers and helping them to improve (or, if that doesn't work, to exit the profession). I look forward to reading the

forthcoming MET reports that tell us more about effective observations; Gates has convened some of the best minds in our country to tackle the issue. I can only hope that the effectiveness of rigorous classroom observations is not judged against the shoddy "truth" of value-added measurements.

Categories: Evaluation