EDUCATION WEEK TEACHER
Published Online: August 27, 2014

FIRST PERSON

Education Is Not 'Moneyball': Why Teachers Can't Trust Value-Added Evaluations Yet

By William Eger

Back to Story

Statistically speaking, public education today is a bit like baseball in the 1990s: The mechanism for evaluating talent is broken. In the 90s, baseball teams, relying on scouts and counting stats like home runs and RBIs, often misidentified average players as stars. Today, teacher-evaluation systems think **everyone is satisfactory**, failing to identify any true MVPs.

In response, school leaders and policymakers, like the general managers of the past, have experimented with quantitative modeling to better measure outcomes. Yet there are reasons to think that approach, arguably successful in baseball, won't help identify good (or bad) teachers.

About 20 years ago, "sabermetricians" revolutionized baseball by analyzing hitters not in terms of isolated stats like home runs or batting average but in terms of their overall value. To try and define value they used regression analysis to create a metric called "Value Over Replacement Player" (VORP), the predecessor of the now ubiquitous Wins Above Replacement (WAR).

The idea behind VORP and WAR is **simple**: The objective is to calculate how many "wins" a batter will generate above a hypothetical standardized backup. For example, this season, catcher Buster Posey has provided 3.5 WAR for the San Francisco Giants, implying that if the Giants had started a typical backup in lieu of Posey, the Giants would likely have won about four fewer games.

While sabermetricians were perfecting VORP calculations, education researchers were rethinking teacher performance. By employing regression analysis in a similar manner, researchers were able to predict a student's expected growth on end-of-year state tests. The difference between the prediction and the student's actual performance would be the "value" that a teacher either added or subtracted during the year.

The resulting systems, known as "Value-Added Models" or VAM, are as conceptually simple as WAR. For example, if a student scores higher than 30 percent her peers on her 9th grade state math test, she would be predicted to score at least that well on her 10th grade state math test. If after taking the test she scored higher than 50 percent of her peers, then her math teacher, the theory goes, is credited with the 20 percent increase.

The Promise of Precise Rankings

The greatest appeal of value-added models is that they differentiate teachers' effectiveness in an elegantly simple way. Instead of going Lake Woebegone on us, and quietly persisting in the logical absurdity that all

"Blindly firing teachers using flawed data without

teachers are above average, value-added systems provide principals and superintendents with a precise percentile ranking of their teachers.

context doesn't give students the best possible teachers."

Value-added models, like WAR, also arguably present a more comprehensive view of a teacher's year-long accomplishment than conventional teacher-evaluation systems. Principals traditionally have relied heavily on a small sample of observations to judge teachers' performance. Last year at my school, which does more observations than most, I was observed for a total of two hours out of roughly a thousand hours of teaching. This is the mathematical equivalent of a baseball scout evaluating a slugger based on a single at-bat. Although a one year-end test is hardly the best measure of a teacher's performance (and the accuracy of value-added models would be improved by incorporating more tests), one test of 130 students is more representative of a teacher's performance than a mere 120 minutes of observed instruction.

Despite these developments in measuring teachers' performances, U.S. Secretary of Education Arne Duncan, **following the lead** of the Gates Foundation, **announced** that he would give states the flexibility to delay the use of assessment results into teacher evaluations. Both the Gates Foundation and the Education Department have been advocates of using value-added models to gauge teacher performance, but my sense is that they are increasingly nervous about accuracy and fairness of the new methodology, especially as schools transition to the Common Core State Standards.

There are definitely grounds for apprehensiveness. Oddly enough, many of the reasons that the similarly structured WAR works in baseball point to reasons why teachers should be skeptical of value-added models.

WAR works because baseball is standardized. All major league baseball players play on the same field, against the same competition with the same rules, and with a sizable sample (162 games). Meanwhile, public schools aren't playing a codified game. They're playing **Calvinball**—the only permanent rule seems to be that you can't play it the same way twice. Within the same school some teachers have SmartBoards while others use blackboards; some have spacious classrooms, while others are in overcrowded closets; some buy their own supplies while others are given all they need. The differences across schools and districts are even larger.

Even when these structural factors are held constant, results can vary wildly. Last year my high school brought in a consulting company to analyze student performance through a value-added model. By analyzing past student performances and overall grades, the company wrote complex models that calculated the expected performance for each of my students on each of my quarterly exams. The average in one class consistently beat expectations while in the next period students consistently performed a full standard deviation below the model's prediction. Same teacher, same material, same classroom, inexplicably different results. If Buster Posey switched teams his WAR wouldn't change—so why is it with teachers?

Subtle Classroom Interactions

At its core, teaching is the summation of subtle teacher-student interactions. These interactions are shaped by both the teacher and the class composition. A class with two innately strong helper students, for example, will likely do better than a class without such students, regardless of the teacher. Strong classes can be derailed by a single broody adolescent in ways that value-added measurement is unable to foresee. Of course excellent teachers can create helper students or inspire a morose teenager. But instead of rewarding exceptional performance, value-added

models will view this teacher as merely on par with another teacher who was luckily assigned more diligent students—or who is "lucky" enough to teach at a school brimming with helpful students and student emotional support structures.

Ultimately value-added models fall short of WAR in terms of effectiveness because teachers, unlike baseball players, don't control their "on-field" performance. In a **statement** cautioning educators about too quickly implementing value-added models, the American Statistical Association reviewed the academic literature and found that teacher quality only explains between 1 and 15 percent of the variations found in VAM.

Like all teachers, I want to know where I stand compared to my peers and how I can improve. I want to know which classes I'm effective in so I can try to understand how to improve my practice. I want my school to use my students' test data to help me grow. I'm a believer in metrics and I want to believe that value-added models will improve education just as sabermetrics improved baseball. But right now there is too much randomness in the subtle human connections between teachers and students for value-added models to have the rigor—and therefore efficacy—of WAR. The impact of a given teacher on performance is too small while the differences between schools are too great to be accurately and consistently quantified. According to a 2012 *New York Times* **story**, for example, New York City's value-added model has resulted in a system "where the margin of error is so wide that that the average confidence interval around each rating spanned 35 percentiles in math and 53 in English," while some teachers are evaluated on a sample "as few as 10 students."

All teachers are not equal and any system that says otherwise is lying. But just because a system mathematically shows differences doesn't make it better. Blindly firing teachers using flawed data without context doesn't give students the best possible teachers. Nor does it help teachers grow. Value added-modes, as they are currently constructed, feel much more like a war on teachers, not a constructive WAR for teachers.

William Eger is a high school math teacher in Philadelphia who writes on issues inside and outside of the classroom. He recently co-authored a ***piece on teacher training*** *for The Atlantic.*

WEB ONLY