Issues with test-based, value-added models of teacher assessment

January 2016
Texas Classroom Teachers Association

Value-added models of teacher assessment use complex statistical techniques to try to estimate teachers’ effects on student achievement. Proponents view them as a primary tool for differentiating teachers based on performance/effectiveness. However, researchers have found that VAM results are unstable over time, subject to bias and imprecision, and rely solely on results from standardized tests that were not designed for that purpose. For these reasons, researchers say they are an invalid and unfair means of teacher assessment.

The weight of the research does not support using VAM. Instead, researchers find that VAM [value-added measurement] and performance-based evaluations of individual teachers remain inaccurate and inappropriate for evaluation, pay, or retention.¹

"Chances are high for misclassification of teacher performance: A teacher classified as ‘adding value’ has a 25 to 50 percent chance of being classified as ‘subtracting value’ the following year(s) and vice versa. This can make the chances that a teacher will be identified as effective the same as the chances of winning a coin toss.”²

- Yield year-to-year fluctuations in teacher ratings that make them unreliable³
- Often invalid because they are often unreliable, and reliability is a qualifying condition for validity⁴
- Can be biased because certain students are almost never randomly assigned to classrooms (like English language learners) and teachers of those students have more difficulties demonstrating value-added than their comparably effective peers
- Fraught with measurement errors caused by missing data, variables that can’t be statistically controlled, differential summer learning losses/gains, etc.
- Rely on standardized assessments not designed for use in VAMs⁵
- Just as student performance is affected by the performance of peers, teacher performance is influenced by the quality of peer teachers and schools.⁶

VAM has outsized impact even if it’s just one of multiple measures

"In a system that places differential weights, but assumes equal validity across measures, even if the student achievement growth component is only a minority share of the weight, it may easily become the primary tipping point in most high stakes personnel decisions.”⁷

- Research hasn’t yet established level of correlation needed between multiple measures to ensure validity.⁸
Even two measures of teacher quality cannot be trusted to determine whether a teacher is “effective” or “ineffective,” especially when one appears to influence or trump the value of the other.\(^9\)

**Even if validity is assumed, VAM has major limitations in its usefulness.**

*“We need to make sure that we have evaluation measures that give us the information we need to improve that group of teachers. A value-added score doesn’t provide much information to teachers or administrators on how to improve a teacher’s performance.”*\(^10\)

- Difficult to integrate into existing evaluation systems in a reliable way\(^11\)
- Limited to identifying top and bottom 5 percent of teachers at best\(^12\)
- Unable to provide information about why or how teachers may improve their practice\(^13\)
- Applicable to only 20 to 35 percent of teachers (for whom student test data from at least two points in time is available)\(^14\)
- Statistical models cannot fully adjust for the fact that some teachers will have a disproportionate number of students who may be exceptionally difficult to teach (students with poorer attendance, who have become homeless, who have severe problems at home, who come into or leave the classroom during the year due to family moves, etc.)\(^15\)

**VAM is nontransparent and often used inappropriately.**

*“... the fact of the matter is that VAMs are so imperfect they should not be used for much of anything unless largely imperfect decisions are desired.”*\(^16\)

- Uses complex statistical formulas not understood by those they affect
- Used inappropriately to make high-stakes decisions about teacher compensation, promotion and termination\(^17\)

**VAM-based polices are not cost-effective for the purpose of raising student achievement.**\(^18\)

*Achievement differences among students are overwhelmingly attributable to factors outside of schools and classrooms (60 percent); teacher impact is only 7 to 10 percent.*\(^19\)

- Broadly, high-stakes implementation of VAM is certainly premature, and likely a significant waste of time and money better spent on problems more pressing and clearly defined\(^20\)
- Measuring value-added is expensive and labor-intensive because it requires testing at every grade level, tracking of data and students, and the capacity to run data through complex statistical formulas.\(^21\)
Examples of state and school district difficulties with VAM

- New York City’s VAM system had a 35 percent error rate in classifying math teachers’ performance and a 53 percent error rate for language arts teachers.\(^{22}\)

- In a June 25, 2012, Open Letter of Concern to Georgia’s governor, a coalition of researchers, educators and teacher evaluation reform advocates expressed alarm about Georgia’s new (VAM) teacher evaluation system. They cited research that supported their recommendation that the state return federal monies related to the project and opt out of Race To the Top (RT3) as Idaho, Indiana, Kansas, Minnesota, Oregon, South Dakota, Virginia, West Virginia and Wyoming had.\(^{23}\)

- In Colorado, many districts have come to recognize that they do not have the capacity to do this work on their own and are concerned about the potential legal challenges that would emerge if personnel decisions are made by an evaluation system that is not valid and reliable.\(^{24}\)

- The Council of Chief State School Officers reported that a survey of the states’ education staff revealed major concerns about state agencies’ lack of capacity to implement VAM evaluation reforms. South Carolina, for example, was identified as a state that simply has “no resources” to design and implement new teacher-evaluation systems.\(^{25}\)

- Two-thirds of respondents to a February 2012 survey of Indiana school superintendents expressed concern about having the capacity to conduct the number of necessary classroom observations necessary to generate VAM evaluations of teachers and for securing training for the personnel responsible for completing them.\(^{26}\)

- The New York teachers’ union filed a lawsuit in 2012 over the details of the state teacher evaluation system. The case has since been resolved, but nonetheless, it has hampered implementation. Delaware delayed its new teacher-evaluation system by one year, giving the state more time to develop student growth measures. Some observers worry about the lack of capacity to execute complicated reform initiatives given tight deadlines. As one Florida reporter said: “Only a handful of districts feel like they’re prepared to do [new teacher evaluations]. Most feel like they’re rushing.”\(^{27}\)

- More than 10 percent of the Ohio districts that originally signed on to receive the federal Race to the Top grant (requiring teacher evaluation based on student test performance) — about 60 districts — bailed out once grants began funneling down to the local level. Some cited concerns that the grants were too small to get the work done. They said, for example, that simply hiring staff to lead the work would eat up most of the money.\(^{28}\)
Resources


9. Id.


18. Id.


20. "What We Know Now (and How It Doesn't Matter),” Posted on August 19, 2013, by P.L. Thomas, Associate Professor of Education at Furman University


25. Id.


28. Id.