EDUCATIONAL LEADERSHIP
November 2012 | Volume **70** | Number **3**
Teacher Evaluation: What's Fair? What's Effective?

Value-Added: The Emperor with No Clothes

Stephen J. Caldas

The trend to use value-added models to rate teachers and principals in New York is psychometrically indefensible.

New York State, among many others, is racing toward the mandated implementation of a teacher and principal evaluation system based in part on something called a value-added model (VAM). In New York, the results of VAM are going to be used to make high-stakes decisions, including who gets tenure and who gets fired.

Value-added models might work well to help us understand in a general sense how some factors influence education outcomes more than others. But they don't work well at all for making fine rating distinctions between individuals—such distinctions were never their purpose. Unfortunately, these models are going to be used as part of a complicated formula to rate teachers and principals in New York State's hastily passed and sloppily implemented Annual Professional Performance Review (APPR) system.

For me—the former state psychometrician of Louisiana who was hired to analyze an equally flawed value-added-type model at the beginning of the "school reform" movement more than two decades ago—what I'm now witnessing in New York and the rest of the United States is deeply disturbing. A grave injustice is being foisted from the top down on educators who are caught up in the most recent crush of reform initiatives.

What Are Value-Added Models—and Do They Work?

Value-added models are mathematically and conceptually complex, which alone argues against their use for evaluation. This complexity is why people can be easily mystified and intimidated by them. The basic idea is that these models are able to statistically determine (or control for) the influence of multiple factors on some outcome measure, like student achievement on standardized tests.

The proposed VAM for New York State's performance review system will supposedly control for variables, including the poverty level, disability status, and English language learner status of a teacher's class of students. This means the "value added" by that teacher on the growth of student achievement on standardized tests over the course of year can be more accurately determined after statistically removing the influence of these outside factors. In a layperson's terms, these models untangle overlapping influences (which we might visualize as the overlapping circles on a Venn diagram). For example, we know that poverty status and ELL status are associated with lower academic outcomes. So in theory, the model should remove this association so that teachers aren't unfairly penalized for student characteristics over which they have no control.

But buyer beware: The validity and reliability of value-added models for rating the effectiveness of teachers, principals, and schools have been roundly rejected by almost the entire psychometric and education research community (Newton, Darling-Hammond, Haertel, & Thomas, 2010).

Much has been written about how New York's new review system was rushed to adoption simply to acquire $700 million of Race to the Top money, with little expert input and almost no consideration for whether the arrangement even made any sense. Some commentators say it will cost districts much more money to implement this pig in a poke than they'll ever get back in special funding. Others, including many of the state's best education administrators, decry the chaos and consternation the initiative has caused local districts scrambling to implement a complex plan (Merchant, 2011).

There's been insufficient attention paid in this debate to the serious methodological problems with value-added models. I'll attempt to show here how this scheme is psychometrically unjustifiable.

But this drama is also about politics and money, which is why I need to explain what happened 22 years ago in the Bayou State. There, I witnessed up-close something similar to what we're seeing play out in New York, although the stakes in the Empire State today are considerably higher. In Louisiana, politics also allowed for the creation of a psychometric monster—but fortunately, politics slew it. I believe there's still hope for New York and other states to slay the value-added models they're currently inviting into their systems.

What I Saw in Louisiana

My introduction to how value-added-type models could be misused came when I accepted a job as a quantitative analyst at the Louisiana Department of Education in fall 1990, just as the state was trying to implement Louisiana's new education reform bill. When the administrators who interviewed me realized that my expertise was in what statisticians call "multiple regression analysis" (the foundation of value-added models), they seemed relieved.

The reason, I soon discovered, was that the conservative Louisiana legislature had passed one of the country's first comprehensive school reform bills (The Children First Act), which included a provision to reward exemplary schools for doing a "better than predicted" job. To implement the lawmakers' vision, the Department of Education had hired handsomely paid outside consultants. These consultants had come up with a value-added-type model to rank schools (although not individual teachers and principals), with the most exemplary getting the greatest monetary rewards and those at the bottom getting no additional money. The proposed model controlled for factors like the poverty rate of the school and other correlates of achievement, much like New York and other states' proposed value-added models for teacher, principal, and school evaluation. The education department wanted a specialist to review this statistical model. My new bosses gave me the consultants' proposed model along with the school data on which it would be based. My charge was to analyze the model using the actual data so we could determine the model's validity. In other words, we wanted to know if it made sense to distribute rewards to particular schools and withhold money from others based on how well this model would rank a group of actual Louisiana schools.

What VAM Can't Do

Before finishing this story, let me explain what value-added models can—and can't—do. In the aggregate, these models can indeed help us better understand how student, classroom, and school characteristics influence education outcomes. However—and this is crucial—when one tries to predict an *individual* student's level of academic achievement (or an individual teacher's class or a specific school's achievement) on the basis of a model that was calculated using data from hundreds or thousands of individuals or schools, there's typically a large margin of error. The error in these models can be huge, which invalidates their use for making accurate individual-level predictions. Even a strong model can have a large margin of error when trying to predict individual-level achievement. Think of it this way: These models try to explain or account for as many of the important factors that influence some aggregate behavior (such as doing well on tests) as they can, but they can never explain 100 percent of any kind of complex human behavior. At best, they might explain 50 or even 60 percent of what influences a behavior like achievement—but they often predict much less than that. The percentage of the behavior that a model doesn't explain can be thought of as the "error" in the model.

Sociologist Carl Bankston and I have published statistical models using value-added-type methodologies and data on more than 33,000 students, controlling for some of the most important correlates of educational achievement (including student and school poverty status, ELL status, family structure, student race, and more). Our models have typically only explained around 20 percent of the total causes of student test scores (Caldas & Bankston, 1997, 1998). Likewise, any state's VAM is going to have much error in its ability to predict teacher effectiveness at the individual level on the basis of aggregate student data, including test scores. And New York is planning to implement a model that hasn't even been adequately piloted, a serious psychometric no-no.[1]

Another way to look at what these value-added models do is to think of them as creating profiles that predict outcomes based on certain criteria. If I were to take any one of the 33,000 students on which Carl Bankston and I based our model explaining academic achievement and identify that individual's race, poverty status, gender, and so on, I could then make a prediction about how well that student should have performed on the basis of these personal factors. But our model, as noted, had a huge margin of error, in part because there are so many other factors that might predict how well a student will score on a test—such as what that kid ate before the test, how much sleep he or she had, and so on. The chances that our prediction of how well the individual student should have performed would match up with how well the student actually performed were slim.

Claiming that our model, or anyone's, is valid for precisely predicting any individual's achievement would be like predicting that a specific person is likely to commit terrorism because he or she belongs to some particular religion—justifying the prediction with the fact that there's a statistically significant correlation between affiliation with that faith and committing a terrorist act. There may be a correlation between that religion and certain actions at the aggregate level, but that can hardly be used to predict individual-level behavior. Yet this is essentially what New York's and other states' VAMs are going to do: use aggregate data to create a model to predict how effective an individual teacher should be on the basis of her students' academic achievement data and a limited set of student characteristics.

New York's new value-added model won't account for many important factors that we know influence how well students perform on tests. As mentioned, the model will control for such

factors as the percentage of a teacher's students who live in poverty, are classified with a disability, and are English language learners. But these factors put together—along with whatever input a teacher may have—probably wouldn't account for even 50 percent of the explanation of how well students perform.

Moreover, there will be much variation from one teacher to another regarding how much relative influence a teacher has on student achievement. Some may have had several low-performing students in their classroom for only one week; others might have had a stable class of high-performing students over the course of a year. This explains in part why credible research shows that teachers who teach students who are lower on the socioeconomic spectrum or disproportionately ELLs are more likely to be rated as ineffective—even after controlling for these factors (Newton et al., 2010).

The Rest of the Story

Let's return to the story of my stint in Louisiana's Department of Education. My colleagues and I determined that the consultants' value-added model had a boatload of error and was consequently a very poor predictor of overall school achievement. When we ranked schools according to this model, the rank ordering didn't even pass the face validity test; people familiar with the record of results and the general quality of various schools in the state could clearly see the rankings didn't reflect long-standing realities and informed perceptions.

But I was a newly minted PhD and I spoke like one. I understood quite well that the model was a psychometric outrage, but I couldn't explain why this was the case in a layperson's terms: I used expressions like *residuals* and *low R-squared*. My colleagues knew that our department would have a hard time explaining to upper echelon administrators, state board members, and state legislators unschooled in statistics why this whole program should be scrapped and the money simply doled out equally to all schools.

Ironically, politics saved the day. The "reform" governor was in trouble during that election year, when we were scheduled to hand out the first School Incentive Awards. Enthusiasm for his school reform initiatives waned, and the legislature cut a lot of School Incentive Program funding. We ended up giving out relatively small checks to the top winners just before a new governor was elected and the School Incentive Program was axed.

Reasoned Resistance

The fact is, value-added models were never intended to be used to accurately predict or rate an individual's performance—and can't do so reliably. Although it may be difficult to explain in layperson's terms why using VAM to rate teacher and principal effectiveness is malarkey, it's not impossible to make this case. I hope I've made a good start.

New York educators who know in their hearts that the current performance review plan is unsound and unjustifiable should realize that there's still hope to change the course. Unfortunately, I've heard few people involved in the New York debate (or elsewhere) discuss these psychometric shortcomings of value-added models, probably because few understand how value-added models work. Terry Orr (2012), an expert in statistics and school effectiveness, has publically noted that APPR is based on evaluation methods that are only now being developed, have never been tested for teacher evaluation, and are of limited use for improving educators' practice. However, she's almost a lone voice. But we can change that. If

there's enough reasoned resistance to this rush to implement a misguided policy, the tide might turn.

New York has always been a leader in education. It should continue the highest traditions of the Empire State and point out the obvious—that using value-added models to rate teachers and principals is folly. This emperor clearly has no clothes.

References

Caldas, S. J., & Bankston, C. L. (1997). The effect of school population socioeconomic status on individual student academic achievement. *Journal of Educational Research, 90*, 269–277.

Caldas, S. J., & Bankston, C. L. (1998). The inequality of separation: Racial composition of schools and academic achievement. *Educational Administration Quarterly, 34*(4), 533–557.

Merchant, R. (2011, January 19). Croton-Harmon, other school districts seek relief from unfunded mandates. *The Journal News*, p. A4.

Newton, X., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-Added Modeling of teacher effectiveness: An exploration of stability across models and contexts. *Educational Policy Analysis Archives, 18*(23). Retrieved from http://epaa.asu.edu/ojs/article/view/810

Orr, M. T. (2012, February). Psychometric problems with the APPR. In A. Dodge (chair). *More than a number: Is the new New York State principal and evaluation system undermining effective teacher learning?* Symposium conducted at the Long Island University, Brookville, New York.

Endnote

[1] For standards regarding appropriate pilot testing of assessments, see American Educational Research Association, National Council for Measurement in Education, & American Psychological Association. (1999). *The standards for educational and psychological testing*. Washington, DC: Author.

Stephen J. Caldas is a professor of education at Manhattanville College in Purchase, New York.