RESEARCH BRIEF

Research Services

Vol. 1502
September 2015

Dr. Aleksandr Shneyderman
Dr. Terry Froman

Using Student Growth to Evaluate Teachers: A Comparison of Three Methods

Results at a Glance

In this brief we compare three methods of teacher evaluation: the State system employing value-added models, a district-level procedure using single-level regression, and a common alternative approach utilizing student growth percentiles. All three methods start by constructing predictions of student test performance based on prior achievement data and student characteristics. At this basic building block level, all three approaches produce virtually identical results. As the methodologies diverge in techniques of aggregation, teacher-level and school-level summary indices begin to separate, but remain remarkably comparable.

In accordance with the federal No Child Left Behind (NCLB) law of 2001, 100% of students were expected to become proficient on state assessments of reading and mathematics by the end of 2013-2014 academic year. Schools that consistently failed to meet the NCLB's Adequate Yearly Progress requirements were subject to penalties. In 2011, the U.S. Department of Education invited each State educational agency (SEA) to request flexibility regarding specific requirements of the NCLB in exchange for "rigorous and comprehensive State-developed plans designed to improve educational outcomes for all students, close achievement gaps, increase equity, and improve the quality of instruction." In order to receive flexibility from the NCLB Adequate Yearly Progress requirements, states had to develop and implement "high-quality teacher and leader evaluation and support systems that are based on multiple measures, including student growth as a significant factor and other measures of professional practice." At the time of the publication of this brief, most states received flexibility waivers. Currently, these states are at different stages in the process of implementing their teacher evaluation and support systems. Many of them use Value-Added Models (VAM) similar to those used in Florida, while others use the Student Growth Percentile (SGP) approach.

In 2014, Florida VAM produced outcomes for teachers of reading in grades 4-10, mathematics in grades 4-8, and algebra in grades 8-9. State law demands that all teachers be evaluated using student growth measures as 50% of the overall evaluation metric (with some exceptions). Thus,

the districts in the State were faced with the necessity of developing their own models of student academic growth. In Miami-Dade, the Covariance Adjustment Model was used for these purposes.

This Research Brief has the following three goals: (1) to describe the workings of the Florida Value-Added Model, the District Covariance Adjustment Model, and the Student Growth Percentile Model, (2) to compare the student- and teacher-level outcomes resulting from these models, and (3) to discuss the use and shortcomings of statistical models designed to evaluate teacher effects on student academic growth.

The Workings of the Models

Florida VAM

The Florida VAM uses the student Florida Comprehensive Assessment Test (FCAT) score for a particular subject area (reading or mathematics) and grade level in a current academic year as an outcome in a multi-level regression model. The scores for two prior year tests in the same subject are used as predictors in the model. In addition, to adjust statistically for student differences among different classrooms, various student and classroom characteristics are used as predictors (covariates) as well. The student-level covariates used are status of a student as an English language learner, gifted status, a set of indicators for any exceptionality a student might have, the difference between the student age and the modal age for the grade level, school attendance, student mobility, and the number of subject-relevant courses the student is enrolled in. The teacher-level covariates are class size, and a measure of dispersion of students' prior year FCAT scores. The result of fitting the model to the data is an equation that can be used to calculate an expected test score based on the values of student- and teacher-level predictors in the model. Then, the actual test score for a student in a current year can be compared with that expected test score. The difference between the actual student test score and the model-based expected score (known as the residual score) is used as a building block in determining measures of teacher and school effects on student test scores. The precision-weighted average of these residuals is regarded as the teacher effect, which is positive if his/her students, on average, performed better than expected, and negative, if the students performed worse than expected.

In addition to providing the teacher effect, the model produces the school component. A positive school component can result from school factors (such as effective leadership, staff training and development, school climate, etc.) or from more effective teachers clustered in high academic growth schools, or a combination of school factors and average teacher effects. The statistical model cannot, in principle, determine the source of the positive (or negative) school component. A State advisory committee decided that a part of the school component

should be credited back to teachers. In accordance with that decision, the model-based estimate of the teacher influence on the students' test score growth, called Teacher VAM Estimate, is calculated as Teacher VAM Estimate = Teacher Effect + 0.5* School Component. The Florida VAM also produces the Standard Error (SE) of the Teacher VAM Estimate, a measure of statistical uncertainty in that result. Details about the Florida VAM can be found in the Technical Report here: http://www.fldoe.org/teaching/performance-evaluation/student-growth.stml.

District Covariance Adjustment Model

The District Model is similar to the Florida VAM in that it also uses the current year score as an outcome in a multiple linear regression, in which the prior year test score and several student demographic characteristics are used as predictors/covariates. The major difference between the District Model and the Florida VAM is that the former is a single-level model, whereas the Florida VAM is a multi-level model, which reflects the actual data structure of students nested within classrooms nested within schools. Another difference is that the District Model uses one prior test score as a covariate and a simplified set of student demographic predictors: ELL status, gifted status, exceptionality status, and a relative age indicator as a proxy for a student having been retained in grade at least once in prior school years.

The District Model produces an expected score, which can be used to calculate student residuals. Aggregated residuals are used as the indicators of teacher influence on student test score growth. The standard errors of these mean residuals are calculated as the ratios of student-level standard deviation of residuals to the square root of the number of student results for a teacher. The details on the workings of the model are available in the Technical Report here: oada.dadeschools.net/VAM%20Information/DistrictModelsTeachEval2013_14RB1404.pdf.

Student Growth Percentile Model

The Student Growth Percentile Model was originally conceived as a descriptive model. For each student it produces a percentile standing of the current year score within a distribution of scores of the student's "academic peers". Academic peers are usually defined as those who had a similar prior year or several prior years' scores. If a student growth percentile is higher than the 50th, it means that the student's test score grew more than that of an average academic peer. For a teacher, students' growth percentiles could be aggregated to a median growth percentile, and that statistic used as a descriptive measure of how much the teacher's students grew academically compared to the typical growth shown by their academic peers.

A typical SGP model is implemented via a statistical technique called quantile regression. This technique allows the use of many predictors including a student's prior year test scores and

demographic characteristics. However, in practice, it is often used with prior year score as the sole predictor. Although the SGP model is generally characterized as descriptive, not causal, in practice its results are used as measures of teachers' effectiveness in teacher evaluation systems throughout the nation in which a particular portion of the overall evaluation is based on students' median growth percentiles.

Comparison of the Model Results

We carried out the comparisons between the outcomes from the three models at three different levels of the hierarchy: student, teacher, and school. In addition, we investigated a relationship between teacher effectiveness estimates resulting from the Florida VAM and the District Covariance Adjustment Model in accountability systems that use both the point estimates and their standard errors.

Student-Level Comparisons

The student residuals are the building blocks for constructing teacher effects in both the Florida VAM and the District Covariance Adjustment Model. In SGP models, student growth percentiles are employed to construct measures of teacher effectiveness. Consequently, we first compared those three fundamental student-level outcomes. To enable meaningful comparisons, we used the student data files provided by the State's contractor (American Institutes for Research, or AIR) to find the student-level residuals resulting from the Florida VAM and to construct the District Covariance Adjustment Model. Once a model was constructed for each grade level and subject area, we calculated the student residuals.

In addition, we ran a quantile regression with the same set of predictors as in the District model, and determined student growth percentiles. We then found correlations among student-level State VAM residuals, District Model residuals, and student growth percentiles. These correlations are shown in the table below.

It can be seen that the values of the correlation coefficients are remarkably high. In reading, they range from .862 to .980 with a median value of .938. In mathematics, they range from .849 to .984 with a median of .938. This indicates that the student-level outcomes from the three models, used as fundamental building blocks to construct measures of teacher and school effectiveness regarding student academic growth, are in very high agreement.

Student-Level Correlations

		Reading			Mathematics		
		State VAM	District Model	SGP	State VAM	District Model	SGP
Grade 4	State VAM	1			1		
	District Model	.980	1		.984	1	
	SGP	.939	.958	1	.945	.960	1
Grade 5	State VAM	1			1		
	District Model	.942	1		.931	1	
	SGP	.901	.957	1	.888	.959	1
Grade 6	State VAM	1			1		
	District Model	.929	1		.961	1	
	SGP	.886	.957	1	.899	.938	1
Grade 7	State VAM	1			1		
	District Model	.938	1		.919	1	
	SGP	.896	.958	1	.859	.938	1
Grade 8	State VAM	1			1		
	District Model	.938	1		.901	1	
	SGP	.895	.959	1	.849	.948	1
Grade 9	State VAM	1					
	District Model	.913	1		N/A		
	SGP	.862	.952	1			
Grade 10	State VAM	1					
	District Model	.931	1		N/A		
	SGP	.884	.954	1			

Teacher- and School-Level Comparisons

To enable the teacher- and school-level comparisons of outcomes between the three models, we calculated the teacher- and school-mean residuals based on the District Covariance-Adjustment Model. Similarly, we found the teacher- and school-level median growth percentiles resulting from the SGP model. Subsequently, we defined the District Model "School Component" as the difference between the school mean residual and the grand mean residual (which was zero), and the "Teacher Effect" as the difference between the teacher mean residual and the school mean residual. Similarly, we defined the Student Growth Percentile Model "School Component" as the difference between the grand median SGP (which was 0.5) and the school median SGP, and the "Teacher Effect" as the difference between the Teacher median SGP and the school median SGP.

We then found correlations between the "Teacher Effects" and "School Components" from the three models for each grade level and subject area. Because teacher- and school outcomes that are based on the small number of student results are likely to be unreliable, we included only

those school and teacher results where the model outcomes were based on at least 10 student results. (We used the same restriction when computing correlation coefficients presented later in this document.) The summary information on these correlations is shown in the table below. Each cell contains the minimum, median, and maximum values of correlation coefficients across the grade levels.

Teacher- and School-Level Correlations from Three Models

		Reading			Mathematics		
		State with District	State with SGP	District with SGP	State with District	State with SGP	District with SGP
Teacher Effects	Minimum	.666	.551	.810	.616	.591	.849
	Median	.698	.592	.851	.792	.713	.866
	Maximum	.834	.718	.874	.882	.799	.934
School Components	Minimum	.687	.681	.917	.741	.674	.921
	Median	.820	.778	.949	.875	.806	.930
	Maximum	.933	.898	.963	.948	.939	.971

Note: the values of non-parametric correlation coefficients (Spearman's rho), which might have been more appropriate due to the non-linear nature of SGPs, were only somewhat higher than those shown in the table above for the Pearson's *r*. We decided to use the values of *r* here and in the following text because of its familiarity to general audiences.

Overall, the values of these correlation coefficients indicate a strong to a very strong positive relationship between the estimates of teacher effects and estimates of school components coming from the three different models.

Because the State used the Teacher VAM Estimates, described previously as the teacher effects plus one-half of the school effects, as measures of teacher effectiveness, it was important to construct analogous measures based on the District Model and SGP model. We calculated these and then found correlations between these estimates coming from the three models. The table below shows the values of the correlation coefficients.

The values of these correlation coefficients are high; in reading they range from .638 to .913 with the median value of .800, while in mathematics they range from .642 to .953 with the median value of .874, indicating a strong relationship between the measures of teacher effectiveness calculated from the three different models.

Teacher-Level Correlations between Estimates of Teacher Effectiveness from Three Models

		Reading			Mathematics		
		State VAM	District Model	SGP	State VAM	District Model	SGP
Grade 4	State VAM	1			1		
	District Model	.913	1		.953	1	
	SGP	.822	.894	1	.902	.949	1
Grade 5	State VAM	1			1		
	District Model	.875	1		.939	1	
	SGP	.800	.883	1	.873	.898	1
Grade 6	State VAM	1			1		
	District Model	.794	1		.898	1	
	SGP	.691	.851	1	.856	.901	1
Grade 7	State VAM	1			1		
	District Model	.766	1		.703	1	
	SGP	.673	.872	1	.675	.864	1
Grade 8	State VAM	1			1		
	District Model	.764	1		.668	1	
	SGP	.685	.886	1	.642	.874	1
Grade 9	State VAM	1					
	District Model	.750	1		N/A		
	SGP	.650	.894	1			
Grade 10	State VAM	1					
	District Model	.735	1		N/A		
	SGP	.638	.841	1			

Teacher-Level Comparisons based on Ratios of Estimates of Teacher Effectiveness to their Standard Errors

It is likely that many teacher accountability systems use both the point estimates of teacher effectiveness and their associated standard errors to assign effectiveness ratings to affected teachers. (Different states may use a different number of effectiveness categories. In Florida, four effectiveness categories of Unsatisfactory, Developing/Needs Improvement, Effective, and Highly Effective are used.) One reasonable way of doing this would be to calculate how many standard errors a particular point estimate is away from the mean of all such estimates. Then, if that distance in either a positive or negative direction exceeds a particular threshold value, a specific rating would be assigned. This process is equivalent to finding the ratio of the difference between a teacher VAM estimate minus the mean of all such estimates to the standard error of that estimate, and then assigning a rating based on the value of that ratio.

Since the subtraction of a constant has no effect on the correlation between two variables, we investigated the relationship between the rating from the State and District models by simply

computing the ratios of teacher VAM estimates to their standard errors based on the two models and then correlating these ratios. The values of the correlation coefficients between these two ratios are shown in the table below.

Teacher-Level Correlations between Teacher Effectiveness Ratios from Two Models

	Reading	Mathematics
Grade 4	.907	.958
Grade 5	.879	.939
Grade 6	.802	.907
Grade 7	.772	.756
Grade 8	.774	.694
Grade 9	.722	NA
Grade 10	.715	NA

The values of correlation coefficients shown in the table above indicate a strong relationship between the teacher effectiveness ratios coming from the State and District models. Because of the strength of this relationship it is likely that categorizations based on the State vs. District models would agree to a rather large extent.

Discussion

We have presented comparisons among three statistical models commonly used in teacher evaluation: the Florida Value-Added (multilevel regression) Model, the District Covariance Adjustment (single level regression) Model, and the Student Growth Percentile (quantile regression) Model. It should be noted that all three of these models attempt to estimate the effect of a teacher on student learning insofar as the student test scores are concerned. Despite differences in statistical approaches, these three models turn out to be surprisingly similar to the extent of being virtually interchangeable in results at the primary building block level of student outcomes (residuals for the State and District models or student growth percentiles for the Student Growth Percentile model). Perhaps this shouldn't surprise us too much considering that they all are based on the same set of data, and the three models use similar sets of predictors including the strong and essential factor of previous performance in the prediction of student results. Since the student outcomes are so similar, aggregations of these results to the teacher and school levels are bound to be highly correlated across models, and the results of our investigations demonstrate that fact. Even the results of analyses involving the ratios of teacher-level results to their standard errors are well aligned for the State and District models.

While the teacher- and school-level outcomes from the three models are highly correlated, they are not identical. Since the models diverge considerably in their methods of analysis and aggregation, differences begin to emerge in teacher and school summary indices, and the results are likely to differ to some degree in their final assignment of teacher effectiveness categories for at least some teachers.

In terms of statistical sophistication, the Florida VAM is the most sophisticated of the three models we examined. However, there is no evidence that greater sophistication leads to greater correctness in estimation of the "true" teacher effects on student learning or even on student test scores. Since there can be no outside gold-standard for measuring teacher effectiveness, it is not known which model produces results that are more valid than those coming from other models.

The fact is, all three of these models are regrettably inadequate when it comes to measuring teacher effectiveness or even its narrower facet of teacher effect on student assessment results. Their fundamental limitation is their inability to control for nuisance factors in full. Because students are not assigned randomly to teachers and teachers are not randomly assigned to schools, extraneous variables associated with student sorting confuse the issue and confound the effect differences. Each method tries, in its own way, to isolate these nuisance factors and control for their influence on student performance. Unfortunately, these extraneous influences can never be fully identified, measured, and statistically compensated for to provide a sufficiently uncontaminated teacher effectiveness indicator.

Therefore, we cannot be certain that the estimates of teacher effects coming from any of these three models truly reflect only the teacher's contribution to the growth in student test scores and no other influences. The list of potential other influences include socioeconomic effects (since in Florida, by law, the socioeconomic status cannot be used as a covariate in a statistical model), parental effects, tutor effects, student health effects, neighborhood effects, etc. Certainly, the reader can think of some other factors that influence student learning that were not taken into account by the statistical models. The point is, a true identification of teacher effects on student learning would require a random assignment of students to schools and classrooms, sufficiently large sample sizes, and a strict control of any factor influencing student learning outcomes; that is, it would require an experimental study in a laboratory-like setting. We are very skeptical as to the ability of any statistical model, however sophisticated, to uncover the true or even serviceable teacher effect in the absence of these conditions.

Policymakers in different states selected different models of using student test scores to evaluate the effectiveness of teachers. Given the similarity of the outcomes produced by the three models on the one hand and the shortcomings of the three approaches one the other, choosing among them is largely a matter of settling for the least offensive. One thing is clear:

the policymakers should not fool themselves into thinking that the estimates of teacher effectiveness produced by any of the current models fully reflect the “true” effectiveness of their teachers.