Comparing Teacher Rankings
The Relationship Between Student Gains and Value Added Rankings

Background

In the fall of 2011, Miami-Dade County Public Schools became one of the first districts to award performance pay bonuses to teachers as part of Race to the Top. Part of the bonus distribution was determined by calculating effectiveness rankings for reading and mathematics teachers. These rankings were based on the percentage of students who made FCAT gains as defined by the Florida school grading system.

At approximately the same time, the State produced its preliminary set of value-added model (VAM) estimates of teacher effectiveness. Annual VAM estimates will be an integral part of the new teacher evaluation system that must be created in accordance with the State’s Student Success Act. Employing a sophisticated hierarchical regression technique, each reading and mathematics teacher was given a value-added rating that took into consideration past student performance and a variety of demographic control variables.

The goal of this paper is to assess how comparable teachers’ rankings are from the two different methods: one based on the simple percentage of student gains and the other based on the State’s value-added model.

Methodology

The methodology for computing learning gains was established by the Florida State Grading System. In the school grades learning model, students could demonstrate learning gains by maintaining a proficient-level score; by increasing their score by one or more achievement levels; or, for students who maintain an FCAT Level 1 or 2, by demonstrating more than a year’s growth on the FCAT developmental scale score. The percentages of students making gains were first found for each teacher’s classes within each of the 3-5, 6-8, or 9-10 grade clusters. These percentages were converted to percentile standings within the grade level clusters. Finally, the weighted averages of the percentile standings across the last three years were found using the numbers of students in each grade level cluster as weights. Based on these weighted averages, teachers who were in the top 10 percent of all reading or mathematics teachers were awarded performance pay.

The second method was based on the State’s Value-Added Model (VAM). The VAM teacher estimates were provided by the State separately for reading and mathematics, and separately for each grade level and each of the three academic years: 2008-09, 2009-10, or 2010-11. These estimates, each based on three years of data, were expressed in terms of the FCAT developmental scale scores. To calculate
teachers’ rankings, these VAM teacher estimates, were first standardized within each grade level and the year of testing. Then, these standardized outcomes were aggregated to the teacher’s level using the number of students as weights. Finally, the aggregated values were converted to percentile ranks.

Correlations

Once the percentile ranks were found using both methods, they were compared for those teachers who had a rank from each method. In reading, there were 2,992 such teachers; in mathematics, there were 2,292 teachers. The overall correlation coefficient between the rankings for both methods was .545 for reading and .617 for mathematics, both positive and statistically significant at the .01 level.

Graphs

It is important to note that the rankings based on gains were most important for the top 10 percent of teachers, since those teachers received individual-based performance pay awards in 2011. Below are the scatter plots based on those teachers with percentile standings in the upper 10 percent of all rankings.

![Scatter plots for reading and mathematics rankings.](image)

It can be seen that most of the data points in both scatter plots are located along the upper horizontal bands of the graph. This indicates that the teachers eligible to receive performance pay awards based on the gains methodology had roughly equivalent VAM ranks in the 70-100 percentile range.

Cross-tabulation

To further investigate the degree of agreement between the ranks from the two systems, the ranks were dichotomized using the 90th percentile as the cutoff point, and the resulting dichotomous ratings were cross-tabulated. A table below presents the results of this cross-tabulation.

<table>
<thead>
<tr>
<th>Gains Rankings</th>
<th>VAM Rankings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading/Mathematics</td>
<td></td>
</tr>
<tr>
<td>Lower 90%</td>
<td>Upper 10%</td>
</tr>
<tr>
<td>2415</td>
<td>258</td>
</tr>
<tr>
<td>210</td>
<td>109</td>
</tr>
</tbody>
</table>

In reading, 2524 out of 2992 teachers (about 84 percent) were categorized in the same way by both rating methods. In mathematics, this agreement percentage was 87 percent. Given that the rating methods are based on different procedures, these are moderately high values. These
percentages were not corrected for the chance agreement. Correcting for chance resulted in Cohen’s Kappa statistic values of .23 for reading and .33 for mathematics. Generally, these values are considered as indicating a “fair” degree of agreement.

Conclusions

Each of the two ranking methods has its advantages and disadvantages. The method using student gains is relatively simple to carry out, and the results are well understood by various stakeholders as it uses the State’s definition of what constitutes a gain. On the other hand, this method does not take into account many differences existing among students taught by different teachers.

By contrast, the State-provided VAM outcomes adjust teacher estimates not only for the 1-2 previous years’ scores, but also for many student characteristics, including English Language Learner and Special Education status. Inclusion of these “fairness” variables is an advantage of the VAM-based method. Then again, the State’s VAM method requires an application of the sophisticated statistical techniques that are not well understood by various stakeholders.

Of course, there is no external standard to determine the validity of either approach. However, their general agreement as presented in this paper at least provides mutual support for their relevance to teacher effectiveness as inferred from student test scores. Future performance pay practices can reasonably rely on the use of either of these methods.