Should Teachers be Evaluated by Student Performance on Standardized Tests?

In January 2015, Governor Cuomo proposed changing the state's teacher evaluation system to increase reliance on measures of student progress on statewide standardized tests, using a so called "Value Added Model." In his 2015 State of the State address, he said:

"Now 38% of high schools students are college ready. 38%. 98.7% of high school teachers are rated effective. How can that be? How can 38% of the students be ready, but 98% of the teachers effective? 31% of third to eight graders are proficient in English, but 99% of the teachers are rated effective. 35% of third to eighth graders are proficient in math but 98% of the math teachers are rated effective. Who are we kidding, my friends? The problem is clear and the solution is clear. We need real, accurate, fair teacher evaluations.

We asked the State Department of Education for their ideas and they gave us their feedback and we accept their recommendation. To reduce the over-testing of students we will eliminate local exams and base 50% of the evaluation on state exams. Second, the other 50% of the evaluations should be limited to independent classroom observations. Teachers may not be rated effective or highly effective unless they are effective in the test and the observation categories. We will stop local score inflation, which is resulted in virtually all teachers being rated by setting scoring bans in the state law."

The proposal was very unpopular with teachers and the unions that represent them. Their opposition led to a boycott of the testing by significant numbers of students in many school

districts. Because of the controversy, the Board of Regents, apparently at the Governor's behest, changed the rules to delay implementation of the rules for four years.

Some critics of the change in direction have argued that the Governor "caved in" to the unions. For example, the New York Post, in an editorial on November 29th, titled "Did the teachers unions just break Andrew Cuomo" said:

For years, Cuomo has been hitting his head against the wall on getting a real state teacher-rating system. Every time he seems to make progress, it's followed by delays, postponements and revisions that ensure nothing meaningful happens.

Now he's reportedly set to give up – abandoning the effort to use student scores on state tests to help judge teacher performance. If so, teachers will be judged subjectively, probably by their own peers. Count on every teacher to rank as just peachy – and incompetents to keep on "teaching."

Just as the teachers unions have demanded all along.

All of us want our children to have good teachers. As a child, while most of my teachers were competent, I was was taught by a few individuals who had no business being teachers. One of my teachers used test questions found in the text books that we used, and permitted us to use the answer keys in the back of the books to find the correct answers. A music teacher gave my class "study hall" on a number of occasions, and put his head on his desk, to sleep. That kind of "teaching" cheats children, by denying them the opportunity to learn.

But, is the use of student progress on standardized tests an accurate way to measure teacher effectiveness? Unfortunately, the answer is no.

There are two basic statistical problems involved in the use of student performance on standardized tests to measure

teacher performance. The first involves the question of whether the students in a given classroom are representative of the entire student population in a school district. Unless students are assigned in a random fashion across the district, variations in student abilities could affect their performance in systematic ways that are unrelated to the effectiveness of teachers. The American Statistical Association points out:

VAM [the test based teacher evaluation method] scores are calculated from classroom-level hererogeneity that is not explained by the background variables in the regression model. Those classroom-level differences may be due in part to other factors that are not included in the model (for example, class size, teaching "high-need" students, or having students who receive extracurricular tutoring). The validity of the VAM scores as a measure of teacher contributions depends on how well the particular regression model adopted adjusts for other factors that might systematically affect, or bias, a teacher's VAM score.

The form of the model may lead to biased VAM scores for some teachers. For example,"gifted" students or those with disabilities may exhibit smaller gains in test scores if the model does not accurately account for their status.

Similarly, the Educational Testing Service, the developers of the College Board exams and others, says:

The fundamental concern is that, if making causal attributions is the goal, then no statistical model, however complex, and no method of analysis, however sophisticated, can fully compensate for the lack of randomization. The problem is that, in the absence of randomization, it is hard to discount alternative explanations for the results that are found. (This explains why many consider randomized experiments the gold standard in scientific work.

Specifically, teacher effects based on statistical estimates may actually represent the combined contributions of many factors in addition to the real teacher contribution we are after. Thus the estimate could be fundamentally off target.

Further, it is usually difficult to determine how off target an estimate is. Clearly, substantial discrepancies would seriously undermine the utility of inferences made on the basis of the analysis.

The second statistical problem stems from the small number of students that most teachers work with. For example, elementary school teachers, with classes of twenty or thirty students, see only a small sample of all the students in a school district. Because of sample variability, those small samples are unlikely to be accurately represent typical students in a school district. Consider the idea of forecasting the result of an election from a sample of 25 voters – the likelihood of an accurate result is small. For that reason, researchers seek large sample sizes to ensure accuracy. ETS describes the problem this way:

With a relatively small number of students contributing to the estimated effect for a particular teacher, the averaging power of randomization can't work for all teachers in a given year. Suppose, for example, that there are a small number of truly disruptive students in a cohort. While all teachers may have an equal chance of finding one (or more) of those students in their class each year, only a few actually will – with potentially deleterious impact on the academic growth of the class in that year. The bottom line is that even if teachers and students come together in more or less random ways, estimated teacher effects can be quite variable from year to year.

Teacher performance is important. As parents, we want our children to have every opportunity to succeed Incompetent teachers can limit that opportunity, so it is important that

the people who teach our children are capable of teaching effectively. Teacher evaluation is an important way in which administrators can help ensure that teachers are competent. But the mindless use of unreliable teacher evaluation methods cannot ensure teacher competency.

So, while it is easy to characterize the Board of Regents decision to postpone implementation of the proposal as a political decision that reflects the power of teacher unions, in fact, the decision reflects the reality that the use of student performance on standardized tests as the primary way to evaluate teachers is not a good way to measure their effectiveness.