Research Detects Bias in Classroom Observations

By Stephen Sawchuk
May 13, 2014

As the rubber hits the road in the implementation of states' revamped teacher-evaluation systems, new research illuminates a troubling source of bias. School principals—when conducting classroom observations—appear to give some teachers an unfair boost based on the students they're assigned to teach, rather than judging them solely on their instructional savvy.

Observers tended to give the best marks to teachers whose students already were high-performing, while those teachers working with academically struggling students were penalized, according to an analysis of thousands of observation scores.

The report, released today by the Brown Center on Education Policy at the Brookings Institution, a Washington think tank, raises a host of new concerns about the nation's evolving systems for grading teachers. And it suggests that, in trying to manage the technical and political challenges posed by test-score-based approaches to evaluation, such as "value added" methods, policymakers may be missing problems in other features of the systems.

"It's very worrisome. It's a huge bias," said Grover J. "Russ" Whitehurst, the director of the Brown Center. "The criticism about value-added is certainly something we need to attend to, but a lot of work has helped reduce or eliminate that bias. None of that's being done for observation scores."

The report recommends that districts try to level the playing field by adjusting teachers' observation scores based on the demographics of the students they instruct.

Among other things, the report also recommends scrapping policies that permit teachers to be judged based on the progress of all students in the school.

District Data

Spurred largely by federal efforts, such as the Race to the Top competition, dozens of states rushed to introduce new teacher-evaluation systems, most based on a combination of test scores and classroom observations.

Both in news stories and in research annals, most of the ink spilled on teacher evaluation has focused on value-added approaches, which estimate each teacher's ability to boost his or her students' standardized-test progress. Observations, in which administrators visit classrooms and rate the quality of a teacher's instruction against a framework, are comparatively understudied.

For their analysis, Mr. Whitehurst and two colleagues examined teacher-evaluation data from four urban school districts ranging in size from 25,000 to 110,000 students. They looked at one to three years of data, analyzing the relationships among the various evaluation components, teachers' overall scores, and the demographics of the students they taught.

The specific weights assigned to each component varied across the districts, but classroom observations counted for at least 40 percent of each teacher's overall score. In most cases, it was more heavily weighed: More than three-quarters of the teacher sample taught in grades or subjects not assessed with standardized tests, and for such teachers, observations are typically weighted more heavily.

Overall, the researchers found that the components' technical properties were consistent with those studied in the massive Measures of Effective Teaching study sponsored by the Seattle-based Bill & Melinda Gates Foundation. But they also discovered some troubling patterns.

For one, the researchers found a strong statistical link between teachers' observation scores and the achievement levels of the students they instructed.

Just 9 percent of teachers of the lowest-achieving students received a top observation score, for example, while 29 percent of such teachers received a ranking in the bottom 20 percent. By contrast, 37 percent of teachers of the highest performing students got a top observation score, and only 11 percent received the lowest score.

News reports indicate that teachers in some districts, including the school system for the District of Columbia, have fretted about similar patterns.

Fix Proffered

The cause of the bias isn't examined in the report, but the authors surmised that some often-measured teaching skills—such as leading a discussion with lots of questioning—may be more difficult with students who are underprepared or not fluent in English.

And the study suggests that the problem may be fixable. The authors applied a handicap of sorts based on student demographics, giving a boost to teachers with many lower-performing students and depressing the scores of those with students who tend to score well. That method would more evenly distribute observation scores across teachers of different groups of students, the paper shows.

Such an idea could be controversial for states and districts to implement because of its assumptions about how much various subgroups of students can progress. But without it, Mr. Whitehurst contends, teacher-evaluation systems may have unintended consequences—such as making working with the neediest students less attractive for teachers.

“Either we have to have observations designed to be immune from this kind of bias, or we have to adjust for it,” Mr. Whitehurst said. “I don’t see any other way out, if we want teachers to teach where we need them to teach, and to be valued for what they do.”

Other observers said that the new research adds yet another question mark to the contested policy push for revamped teacher evaluations.

“I think this is going to be another bad-news story for the supporters of teacher evaluation, and for the [Obama] administration,” said Michael Petrilli, the executive vice president of the Thomas B. Fordham Institute, a Washington think tank. “Rather than trying to find some kind of technocratic solution, we need to get back to common sense—trusting principals to make judgments. If we don’t do that, none of our school-reform efforts are going to work.”

Teachers’ unions have tended to be far more critical of the value-added approach than classroom observations.

Segun Eubanks, the director of teacher quality for the 3 million-member National Education Association, said that, on the one hand, the new analysis from Brookings confirms a general sense among teachers that they’re put at a disadvantage by choosing to work with the most at-risk students.

On the other hand, the notion of adjusting observation scores seems premature without further research, he said.

“My first instinct would be to help to put observational data into a context-specific realm. You need to train folks to see what teaching performance looks like when you’re teaching students who have low achievement,” Mr. Eubanks said. “The look-fors are different; the way the standards are applied are different. We have to find ways to do that before we start going for handicapping.”

Schoolwide Gauge Questioned

The Brookings report also examines a host of other aspects of the newly designed teacher-evaluation systems.

Consistent with the Gates research findings, it suggests that more observations improve the accuracy of the systems, and that outside observers tend to give ratings that are more predictive of teaching quality than principals.

Also, the report takes aim at evaluation systems that use a “schoolwide” value-added measure, in which all teachers are judged in part on the progress of the school as a whole. Such a policy, the report notes, tends to bring down the scores of even good teachers in schools with lots of low-achieving students—and to inflate the scores of weaker teachers who were in high-performing schools.

“It creates a system that is demonstrably and palpably unfair to teachers, given that they have little control over the performance of the whole school,” the report states.

Vol. 33, Issue 32, Pages 1, 10-11

Published in Print: May 21, 2014, as Bias Detected in Classroom Observations