The New York Times Reprints

December 26, 2010

Hurdles Emerge in Rising Effort to Rate Teachers

By SHARON OTTERMAN

For the past three years, Katie Ward and Melanie McIver have worked as a team at Public School 321 in Park Slope, Brooklyn, teaching a fourth-grade class. But on the reports that rank the city's teachers based on their students’ standardized test scores, Ms. Ward’s name is nowhere to be found.

“I feel as though I don’t exist,” she said last Monday, looking up from playing a vocabulary game with her students.

Down the hall, Deirdre Corcoran, a fifth-grade teacher, received a ranking for a year when she was out on child-care leave. In three other classrooms at this highly ranked school, fourth-grade teachers were ranked among the worst in the city at teaching math, even though their students’ average score on the state math exam was close to four, the highest score.

“If I thought they gave accurate information, I would take them more seriously,” the principal of P.S. 321, Elizabeth Phillips, said about the rankings. “But some of my best teachers have the absolute worst scores,” she said, adding that she had based her assessment of those teachers on “classroom observations, talking to the children and the number of parents begging me to put their kids in their classes.”

It is becoming common practice nationally to rank teachers for their effectiveness, or value added, a measure that is defined as how much a teacher contributes to student progress on standardized tests. The practice was strongly supported by President Obama’s education grant competition, Race to the Top, and large school districts, including those in Houston, Dallas, Denver, Minneapolis and Washington, have begun to use a form of it.

But the experience in New York City shows just how difficult it can be to come up with a system that gains acceptance as being fair and accurate. The rankings are based on an algorithm that few other than statisticians can understand, and on tests that the state has said were too narrow and predictable. Most teachers' scores fall somewhere in a wide range, with perfection statistically impossible. And the system has also suffered from the everyday problems inherent in managing busy urban schools, like the challenge of using old files and computer databases to ensure that the right teachers are matched to the right students.

All of this was not as important when the teacher rankings were an internal matter that principals could choose to heed or ignore. City officials had pledged to the teachers' union that the rankings would not be used in the evaluation of teachers and that they would resist releasing them to the public.

But over the past several months, the system of teacher rankings has been catapulted to one of the most contentious issues facing the city's 80,000-member teaching force. A new state law, passed this year to help New York win Race to the Top money, pledges that by 2013, 25 percent of a teacher's evaluation be based on a value-added system. The city has begun urging principals to consider rankings when deciding whether to grant tenure. And the city now supports the release of the data to the 12 media organizations, including The New York Times, that have requested it.

The departing schools chancellor, Joel I. Klein, defended the release of the rankings in an e-mail to school staff members, acknowledging that they had limitations but calling them "the fairest systemwide way we have to assess the real impact of teachers on student learning."

"For too long," Mr. Klein wrote, "parents have been left out of the equation, left to pray each year that the teacher greeting their children on the first day of school is truly great, but with no real knowledge of whether that is the case, and with no recourse if it's not."

But the United Federation of Teachers, the city's teachers' union, has sued to keep names in the rankings private, arguing that the data is flawed and would result in unnecessary harm to the reputation of teachers. The matter is now before Justice Cynthia Kern of State Supreme Court in Manhattan.

New York City began ranking teachers in the 2007-8 school year as part of a pilot project intended to improve classroom instruction. The project, which cost $1.3 million, with an additional $2.3 million budgeted over the next 18 months, was expanded in the 2008-9 school year to give rankings to more than 12,000 fourth- through eighth-grade teachers.

In New York City, a curve dictates that each year 50 percent of teachers will receive “average” rankings, 20 percent each “above average” and “below average,” and 5 percent each “high” and “low.” Teachers get separate rankings for math and English.

In support of the model, Douglas Staiger, an economics professor at Dartmouth College, cites research showing that if a teacher receives a high-performing score one year, there is a modest likelihood that he or she will receive a high-performing score the following year. The correlation is about 0.3, he said, with 1 being perfect, and 0 being no correlation. This means that about one-third of teachers ranked in the top 25 percent would appear among the top quarter of teachers the next year.

While that year-to-year link may seem low, in the budding and messy exercise of trying to quantify what makes students learn, it is one of the strongest predictors of future student performance, along with the reduction of class size. That means that, on average, students placed for a year with a high-value-added teacher will do better than those placed with a low-value-added teacher. Dr. Staiger placed the improvement at about three percentile points on a typical standardized test.

“This information is useful but has to be used with caution,” he said. “It’s that middle ground. It’s not useless, but it’s not perfect.”

Yet a promising correlation for groups of teachers on the average may be of little help to the individual teacher, who faces, at least for the near future, a notable chance of being misjudged by the ranking system, particularly when it is based on only a few years of scores. One national study published in July by Mathematica Policy Research, conducted for the Department of Education, found that with one year of data, a teacher was likely to be misclassified 35 percent of the time. With three years of data, the error rate was 25 percent. With 10 years of data, the error rate dropped to 12 percent. The city has four years of data.

The most extensive independent study of New York's teacher rankings found similar variability. In math, about a quarter of the lowest-ranking teachers in 2007 ended up among the highest-ranking teachers in 2008. In English, most low performers in 2007 did not remain low performers the next year, said Sean P. Corcoran, the author of the study for the Annenberg Institute for School Reform, who is an assistant professor of educational economics at New York University.

The high margin of error for most scores, something the city refers to as the confidence interval, is another source of uncertainty, Dr. Corcoran said. In math, judging a teacher over three years, the average confidence interval was 34 points, meaning a city teacher who was ranked in the 63rd percentile actually had a score anywhere between the 46th and 80th percentiles, with the 63rd percentile as the most likely score. Even then, the ranking is only 95 percent certain. The result is that half of the city's ranked teachers were statistically indistinguishable.

“The issue is when you try to take this down to the level of the individual teacher, you get very little information,” Dr. Corcoran said. The only rankings that people can put any stock in, he said, are those that are “consistently high or low,” but even those are imperfect.

“So if you have a teacher consistently in the top 10 percent,” he said, “the chances are she is doing something right, and a teacher in the bottom 10 percent needs some attention. Everything in between, you really know nothing.”

In New York, the rankings face an additional set of issues. The state tests on which they were based became, over time, too predictable and easy to pass, and this summer the state began to toughen standards. Daniel Koretz, a Harvard professor whose research helped persuade the state to toughen standards, said that as a result it was impossible to know whether rising scores in a classroom were due to inappropriate test preparation or gains in real learning. Rankings that include the tougher standards will not be available until the next academic year.

“It would make sense to wait until the problems with the state test are sorted out, because we are going to get it wrong a lot of the time,” Dr. Koretz said.

City officials defended using the state tests as a basis for the rankings, saying that they remained predictive of other outcomes, like graduation rates. Echoing Dr. Corcoran, the officials said they were most interested in identifying teachers at the extremes. “We have read the studies on it, and it is the best quantitative method that we have,” said John White, a deputy chancellor. “When used in concert with other pieces of information, it can help us judge teacher effectiveness.”

Beyond the formulas and tests, individual errors — like the one that led Ms. Ward to be left out altogether — have generated controversy. The teachers’ union claims that it has found at least 200 such errors, including teachers’ getting rankings for subjects they did not teach (sometimes they did well, sometimes poorly). Mr. White would not provide an estimate for the error rate, but noted that principals had 18 months to correct mistakes in class lists, starting from when the scores were first distributed.

Mr. White said on Tuesday that before the next round of rankings was released, teachers would be able to review class lists to verify which students they taught, a practice that generally did not happen in the past. Douglas N. Harris, an economist affiliated with the center at the University of Wisconsin that produces the city’s rankings, called the science behind them promising, and said that they had jump-started a wider effort to come up with better measures of teacher performance, which was long overdue.

But Dr. Harris urged caution in reading too much into the early crop of rankings, and added, “As a general rule, you should be worried when the people who are producing something are the ones who are most worried about using it.”