EDUCATIONAL LEADERSHIP

November 2012 | Volume **70** | Number **3**
Teacher Evaluation: What's Fair? What's Effective? Pages 54-57

The Fuzzy Scarlet Letter

Aaron M. Pallas

The public release of teacher evaluation scores is unfair and misleading—and it provides little useful information for parents.

Critics of the public release of teacher evaluation scores sometimes liken these ratings to the scarlet letter worn by Hester Prynne in Nathaniel Hawthorne's classic novel. The comparison is apt. But public school teachers who are subjected to public shaming because of their students' test scores can rarely expect the opportunities for redemption offered to Prynne, whose humility and good works over time changed the meaning of her scarlet A from "Adulteress" to "Able."

U.S. political and economic leaders tell us that serious problems require bold action. In the realm of public education, this has meant a rapid expansion of systems intended to hold schools and teachers accountable for student performance. Such accountability has been applied to schools for 10 years under No Child Left Behind. But there is still considerable debate over whether individual teachers should face public accountability for the results of their evaluations. After all, personnel evaluations in most sectors of the economy are viewed as a private matter between employer and employee. Should it be any different for teachers?

Teachers as a Commodity

The debate between transparency and privacy is framed by the multiple purposes that teacher evaluations are intended to serve—purposes that are often in tension with one another. Some policymakers and practitioners emphasize the use of teacher evaluations for *selection*—to weed out ineffective teachers and perhaps identify the best ones for rewards, such as merit pay. Others view teacher evaluation as a tool for *direction*, pointing teachers toward aspects of their classroom practice that they can improve.

But now we are witnessing the emergence of another, more insidious view of teacher evaluation—one that frames teacher quality as a commodity that can be exchanged in the marketplace. In this view, evaluations signal to parents that some teachers are better than others. Who wouldn't prefer their child to have a teacher judged "highly effective" over one judged "ineffective"?

Such ratings can kick off a market-based competition in which parents try to bargain to acquire the highest-quality teacher for their own children. Those who approve of such a use of teacher evaluations argue that parents have the right to information that will enable them to exercise choice in the marketplace of schools and teachers. This logic, casting parents as consumers in a free market, has been central to debates about the public release of teacher value-added scores as well as formal teacher evaluations.

So Far, Parents Aren't Biting

We don't have a great deal of evidence about the mischief that linking teachers' names to their evaluations might cause. To date, the most publicized disclosures of teacher rankings have been unofficial and partial rankings, not official and comprehensive ones. In 2010, the *Los Angeles Times* matched Los Angeles students' test scores to their teachers and published value-added scores produced by a respected economist using a complex statistical model. The newspaper conveyed some information about the margin of error in the scores and provided some technical material about the methods used, but it largely ignored its own caveats in describing particular teachers as more or less effective. The outcry from teachers and others was immediate (Gardner, 2010).

A year later, the *Los Angeles Times* updated its ratings, using the most recent year of test data available. Stung by the initial reaction and critiques, the *Times* revamped its online reports, conveying the uncertainty in the measures more successfully. The second time around, public reaction was muted. In both cases, few parents pressured principals to move their children to the higher-ranking teachers or clamored to escape from the lower-ranking ones.

In New York City, the Department of Education goaded the major media organizations into requesting its Teacher Data

Reports—the school district's internal value-added scores produced for reading and math teachers in grades 4–8 (Hancock, 2011). The United Federation of Teachers, New York City's teachers union, sued to block their release; when the litigation subsided, with the courts ruling that the individually identifiable ratings could be released to the media, the department released ratings from 2007–08, 2008–09, and 2009–10. The local media publicized these ratings widely in the winter of 2012, with the tabloids publishing front-page stories, replete with ambush-style photos, about the teachers identified as the "worst" teachers in New York City (Chapman, Lesser, & Fanelli, 2012; Macintosh, 2012; Roberts, 2012).

There's little question that the published ratings subjected a great many New York City teachers to shame and ridicule. One of my own students, a teacher in Brooklyn, reported overhearing a student say to one of her colleagues, "I don't have to listen to you! You got a *D*!" And at least one outstanding middle school math teacher decided to leave teaching at least partly due to a Teacher Data Report that ranked her as the worst 8th grade math teacher in New York City (Pallas, 2012).

The political leaders who engineered the public release of the Teacher Data Reports may have misjudged the political fallout. The annual ratings were shown to be highly imprecise, with an average confidence interval of more than 50 percentiles (Corcoran, 2010). And many parents, as well as others in the school community, saw that the ratings didn't align with their own impressions of their children's teachers. As in Los Angeles, there is little evidence so far that New York's publication of unofficial teacher rankings has resulted in new pressures from parents to place their children in the classrooms of particular teachers.

How to Choose?

To illustrate the difficulty parents would have in identifying the "best" teachers on the basis of public ratings, I examined the value-added scores of 2,656 New York City 4th grade teachers who taught reading in the 2009–10 school year. These teachers were spread across 703 elementary schools, with an average of about four teachers per school; 33 schools had just one 4th grade teacher who taught reading, and three schools had 10 or more.

As researchers and scholars alike have noted, these value-added scores are not precise. I found that for more than half of the teachers, the range of plausible scores was so wide that it's unclear whether they were in the top 20 percent of teachers or the bottom 20 percent.

To take this one step further, I compared the scores of the 4th grade reading teachers who were teaching in the same school. This comparison is presumably what many proponents of the public release of teacher evaluations have in mind: empowering parents to choose the best teacher in the school for their child.

First, let's admit that there are several problematic assumptions weighing down the free-market concept here. For example, a value-added measure might purport to describe Ms. Walters's contribution to the achievement of her average student, but if your child isn't average, the value-added measure may not indicate how your child would fare in her classroom.

Second, a value-added score for Ms. Walters will almost certainly pertain to her performance in a prior year, not to the current year or a future year in which your child might be in her classroom. Given what we know about how unstable value-added measures are from one year to the next, perhaps the fine print should mimic that in automobile ads: "Your mileage may vary."

And finally, there's the peculiar notion that parents are actually allowed to choose their child's teacher within a school. In most schools, the mechanisms for assigning students to teachers' classrooms don't provide much room for parental choice because allowing parents to pick the most popular teacher in the school for their child would be neither equitable nor efficient.

With these caveats in mind, how might a parent go about using value-added rankings to pick a teacher for his or her child? Of course, in the 44 elementary schools in New York City with only a single 4th grade reading teacher, there would be no choice involved. In 2009–10, there were 146 elementary schools with two 4th grade reading teachers, and one of the teachers had a higher value-added score than the other in just 26 of them (18 percent). The percentage was similar for the 522 schools with three or more 4th grade reading teachers.

Overall, if we tabulate the nearly 4,800 comparisons among pairs of 4th grade reading teachers teaching in the same New York City elementary school, one teacher in a pair was demonstrably more successful in promoting performance on the state English language arts test *in only 12 percent of the comparisons*. And we would expect teachers to differ 5 percent of the time by chance alone.

Of course, many of these 4th grade teachers who were teaching reading were also teaching mathematics, and their value-added scores on the state mathematics test could be subjected to a similar analysis. When teachers teach multiple subjects in self-contained classrooms, parents don't have the luxury of mixing and matching teachers to optimize their children's achievement. A teacher who is systematically better than another on all of the dimensions that a parent might judge important would be a rarity.

Rigid Boxes

Public education is shifting from a system of perfunctory evaluations, with few consequences for teachers and schools, to evaluations that are used to support high-stakes decisions about teachers' employment prospects.

The problem with making such evaluations public is that virtually all methods of evaluating teachers have both random and nonrandom errors that may mask a teacher's true performance. Accurate classroom observations, for example, require extensive training of observers (more than most school districts now provide); and even assuming expert evaluators, how well a class session goes can vary substantially from day to day.

Student test scores are even more problematic: Even if we really believed that a state's 4th grade math tests covered all of the mathematics that students should learn in that grade (and does anyone really believe this?), any effort to use students' test scores to compare one teacher with another will be imprecise because the comparisons are based on small samples of students and curricular content.

The state of New York has now passed a law requiring its school districts to publicly report teachers' composite effectiveness scores ranging from 0 to 100 as well as their final rating of *highly effective, effective, developing*, or *ineffective* (New York State Education Department, 2012). Variations of this approach have been adopted in many other states and school districts. But a score of 75 out of 100 points on an annual evaluation, or a summary grade of *developing*, may not convey that that 75 could just as easily have been a 64 or an 89, or that the teacher classified as *developing* might actually be *effective*. We can always construct rules and algorithms that will assign someone an unequivocal rating, but without the appropriate context, this rating will appear to be much more precise than it actually is.

In spite of the inherent uncertainty in teacher evaluations, policymakers want to treat the evaluation measures as though they are infallible and use them to place teachers in rigid boxes, labeling them as good teachers or poor teachers. Policymakers and the media treat these labels as definitive, but the raw material being stuffed into the boxes will rarely fit in one box without spilling over into the adjacent ones.

If states and school districts insist on publicizing individual teachers' evaluation scores—slapping a metaphorical scarlet A on some teachers and a stamp of approval on others—the only fair thing to do is to admit that the scarlet letter is fuzzy, a bit out of focus. Anything else is bad fiction.

References

Chapman, B., Lesser, B., & Fanelli, J. (2012, February 24). More than a dozen teachers earned lowest scores on controversial rankings. *New York Daily News*. Retrieved from www.nydailynews.com/news/a-dozen-teachers-earned-lowest-scores-controversial-rankings-article-1.1028113

Corcoran, S. P. (2010). *Can teachers be evaluated by their students' test scores? Should they be? The use of value-added measures of teacher effectiveness in policy and practice*. Providence, RI: Annenberg Institute for School Reform at Brown University.

Gardner, W. (2010, October 1). Suicide of teacher and published rankings. [blog post]. Retrieved from *Walt Gardner's Reality Check*.

Hancock, L. (2011, March/April). Tested: Covering schools in the age of micro-measurement. *Columbia Journalism Review*. Retrieved from www.cjr.org/cover_story/tested.php

Macintosh, J. (2012, February 25). Teachers who got zero ratings. *New York Post*. Retrieved from www.nypost.com

New York State Education Department. (2012). *Guidance on New York State's Annual Professional Performance Review for teachers and principals to implement Education Law 3012-c and the commissioner's regulations*. Albany, NY: Author.

Pallas, A. (2012, May 15). The worst eighth-grade math teacher in New York City. [blog post]. Retrieved from *A sociological eye on education* at http://eyeoned.org/content/the-worst-eighth-grade-math-teacher-in-new-york-city_326

Roberts, G. (2012, February 26). Queens parents demand answers following teacher's low grades. *New York Post*. Retrieved from www.nypost.com

Aaron M. Pallas is a professor of sociology and education in the Department of Education Policy and Social Analysis, Teachers College, Columbia University, New York.