EDUCATIONAL LEADERSHIP

November 2012 | Volume **70** | Number **3**
Teacher Evaluation: What's Fair? What's Effective? Pages 80-81

Use Caution with Value-Added Measures

Bryan Goodwin and Kirsten Miller

When the New York City Department of Education released its Teacher Data Reports in February 2012, Pascale Mauclair found herself in the spotlight—for all the wrong reasons. The *New York Post* dubbed Ms. Mauclair, a 6th grade teacher at highly rated P.S. 11 in Queens, the "city's worst teacher." There was just one problem. It wasn't true.

First, the data were suspect: Of the seven 6th grade teachers in the same school, three received zero percentile scores, an unlikely scenario for a school rated in the 94th percentile of the city's public schools. Next, although Ms. Mauclair taught both math and English language arts, only six of her students had taken the language arts assessment, a number below the allowable reporting sample of 20 students. Her value-added rating was therefore based solely on the results for the 11 students who took the mathematics exam (for which the minimum reporting sample is 10 students). Such a small sample is prone to distortions. Further, her class consisted of immigrant students who were still learning English and who entered her classroom at different times during the year; some students took the exam when they had been in her class for just a few months (Casey, 2012; Clawson, 2012).

Clearly, the numbers didn't tell the whole story. Yet Mauclair, who was regarded by other teachers and administrators in this high-performing school as an excellent teacher, was held up to public criticism by those unaware of the realities of the situation (Casey, 2012). In light of her experience and the similar experiences of other teachers, we should ask what the research says about the accuracy of value-added measures of teacher performance.

Researcher Misgivings

In many ways, the value-added teacher measurement model is still in its infancy, having emerged only in recent years as sophisticated data warehouses made it possible to measure the average growth of an entire class of students over the course of a school year. However, researchers have warned that what seems so simple and straightforward in theory is incredibly complicated in practice. Here are a few of the pitfalls.

Non-teacher effects may cloud the results. Meta-analytic research conducted by Marzano (2000) found that teachers account for only about 13 percent of the variance in student achievement. Student variables (including home environment, student motivation, and prior knowledge) account for 80 percent of the variance. Value-added models don't necessarily isolate teacher effects from these other influences (Braun, 2005).

Data may be inaccurate. In the aftermath of the Pascale Mauclair incident, multiple factual errors surfaced in New York's data. For example, one teacher had data for a year when she was on maternity leave; another teacher taught 4th grade for five years but had no data (Clawson, 2012). Moreover, small samples—for example, classes with only 10 students—can paint inaccurate pictures of teachers because they are subject to statistical fluctuations (Goe, Bell, & Little, 2008).

Student placement in classrooms is not random. For a variety of reasons, schools seldom place students randomly in classrooms. As a result, some teachers find themselves with accelerated learners, whereas others, like Ms. Mauclair, may find themselves with more challenging students. Existing models do not adequately control for this problem of nonrandom assignment (Rothstein, 2008).

Students' previous teachers can create a halo (or pitchfork) effect. Researchers have discerned that the benefits for students of being placed in the classrooms of highly effective teachers can persist for years. As a result, mediocre teachers may benefit from the afterglow of students' exposure to effective teachers. Conversely, researchers have found "little evidence that subsequent effective teachers can offset the effects of ineffective ones" (Sanders & Horn, 1986, p. 247). As a result, the value-added ratings for effective teachers may be diminished because of previous, ineffective teachers.

Teachers' year-to-year scores vary widely. Perhaps one of the most troubling aspects of value-added measures is that the ratings of individual teachers typically vary significantly from year to year (Baker et al., 2010). For example, in one study, 16 percent of teachers who were rated in the top quartile one year had moved to the bottom two quartiles by the next year, and 8 percent of teachers in the bottom quartile had risen to the top quartile a year later (Aaronson, Barrow, & Sander, 2003).

Still Better Than the Alternatives?

In general, the year-to-year correlation between value-added scores lies in the .30 to .40 range (Goldhaber & Hansen, 2010). Although this correlation is not large, researchers at the Brookings Institution note that it is almost identical to the correlation between SAT scores and college grade point average (.35); yet we continue to use SAT scores in making decisions about college admissions "because even though the prediction of success from SAT/ACT scores is modest, it is among the strongest available predictors" (Glazerman et al., 2010, p. 7).

Similarly, more traditional measures of teacher performance have not been tremendously accurate. For example, until recently, many teacher evaluation systems only provided binary ratings: *satisfactory* or *unsatisfactory*, with a full 99 percent of teachers receiving *satisfactory* (Weisberg, Sexton, Mulhern, & Keeling, 2009). Moreover, researchers have found weak correlations between principals' ratings of teacher performance and actual student achievement; in general, principals appear to be fairly accurate in identifying top and bottom performers, but they struggle to differentiate among teachers in the middle (Jacob & Lefgren, 2008).

When faced with imperfect predictors of college success, colleges have learned to use a variety of measures to make decisions about which students to admit. The challenges posed by value-added measurement would suggest that schools take a similar approach. School leaders should heed researchers' consistent warnings against publicly releasing individual teacher ratings or relying heavily on value-added measures to make high-stakes employment decisions. But value-added measures might reasonably be considered as one component of teacher evaluation—when taken with a healthy dose of caution and considered alongside other measures.

References

Aaronson, D., Barrow, L., & Sander, W. (2003). *Teachers and student achievement in the Chicago public high schools*. Chicago: Federal Reserve Bank of Chicago.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., Ravitch, D., et al. (2010). *Problems with the use of student test scores to evaluate teachers*. Washington, DC: Economic Policy Institute.

Braun, H. I. (2005). *Using student progress to evaluate teachers: A primer on value-added models*. Princeton, NJ: Educational Testing Service.

Casey, L. (2012, February 28). The true story of Pascale Mauclair. *Edwize*. Retrieved from www.edwize.org/the-true-story-of-pascale-mauclair

Clawson, L. (2012, March 4). New York City's flawed data fuel right's war on teachers. *Daily Kos*. Retrieved from www.dailykos.com/story/2012/30/04/1069927/-New-York-City-s-flawed-data-fuels-the-right-s-war-on-teachers

Glazerman, S., Loeb, S., Goldhaber, D., Steiger, D., Raudenbush, S., Whitehurst, G. (2010). *Evaluating teachers: The important role of value-added*. New York: Brookings. Retrieved from www.brookings.edu/research/reports/2010/11/17-evaluating-teachers

Goldhaber, D., & Hansen, M. (2010). *Assessing the potential of using value-added estimates of teacher job performance for making tenure decisions* (Working paper 31). Washington, DC: National Center for Analysis of Longitudinal Data in Education Research.

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.

Jacob, B. A., & Lefgren, L. (2008). *Principals as agents: Subjective performance measurement in education* (Faculty research working papers series No. RWP05-040). Cambridge, MA: Harvard University John F. Kennedy School of Government.

Marzano, R. J. (2000). *A new era of school reform: Going where the research takes us*. Aurora, CO: McREL.

Rothstein, J. (2008). *Student sorting and bias in value-added estimation: Selection on observables and unobservables*. Paper presented at the National Conference on Value-Added Modeling, Madison, WI. Retrieved from www.wcer.wisc.edu/news/events/vam%20conference%20final%20papers/studentsorting&bias_jrothstein.pdf

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education, 12*(3), 247–256. Retrieved from www.sas.com/govedu/edu/ed_eval.pdf

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. Brooklyn, NY: New Teacher Project.

Bryan Goodwin is vice president of communications, McREL, Denver, Colorado. He is the author of *Simply Better: Doing What Matters Most to Change the Odds for Student Success* (ASCD, 2011). Kirsten Miller is a lead consultant at McREL.