Stanford University • September 2010

SIEPR policy brief

Stanford Institute for Economic Policy Research

on the web: http://siepr.stanford.edu

# Student Assessment and Teacher Effectiveness: Telling the Whole Story

*By Maria D. Fitzpatrick*

Recently, the *Los Angeles Times* published an explosive article reporting the results of its assessment of more than 6,000 teachers in the Los Angeles Unified School District (Feltch et al. 2010). The *Times'* analysis uses a version of a conventional technique for measuring teacher effectiveness, value-added modeling (VAM). VAM evaluates teachers by comparing the performance of students in their classes on standardized tests given before and after the year spent with the teacher. The article has drawn fire for many reasons—from the quality of the study's design to its singling out of individual teachers. Indeed, one of the most striking features of the article is that the newspaper has promised to provide a value-added score, or measure of teacher effectiveness, for each teacher by name on its website.

Economists generally believe that information is a good thing. Its provision leads to enhanced market functioning. Information asymmetries between interacting parties, like parents and teachers, can lead to market failure. While the *Times'* analysis likely has some flaws, the publication has jump-started a long-needed public discussion about teacher assessment and accountability.

A body of evidence has amassed showing that teachers can have a profound impact on student test score performance from year to year (Rivkin et al. 2005; Rockoff 2004). By providing parents and taxpayers with information about teacher effectiveness they otherwise would not have, the *Times'* analysis has the potential to improve children's education. Those of us involved in education, however, know that doing so will not be easy. There is still a great deal of uncertainty about the correct way to measure and report teacher effectiveness. By using this brief to set forth some of the issues involved, I

*continued on inside...*

## About The Author

**Maria Donovan Fitzpatrick** is a Searle Freedom Trust Postdoctoral Scholar at the Stanford Institute for Economic Policy Research. She received her BA in economics from University of North Carolina Chapel Hill and her MA and PhD degrees from the University of Virginia. Her research interests center on the economics of education, particularly early childhood education. Her latest publication, *Starting School at Four: The Effects of University Pre-Kindergarten on Children's Academic Achievement,* was published in the B.E. Press Journal of Economic Policy and Analysis in 2008.

hope to continue the conversation started by the *LA Times*.

## What If Tests Do Not Tell the Whole Story?

Any particular standardized test is a measure of a specific thing. In general, the focus of standardized tests implemented as the result of No Child Left Behind (NCLB) is to measure "student achievement." This includes the ability of students to perform tasks like reading comprehension or mathematical problem solving that assessment developers have deemed appropriate for a student at a particular grade level. However, what skills and tasks are appropriate and how exactly we measure proficiency is the subject of a sizable amount of research and debate. Yet the rewards and sanctions tied to the tests can be incredibly large.

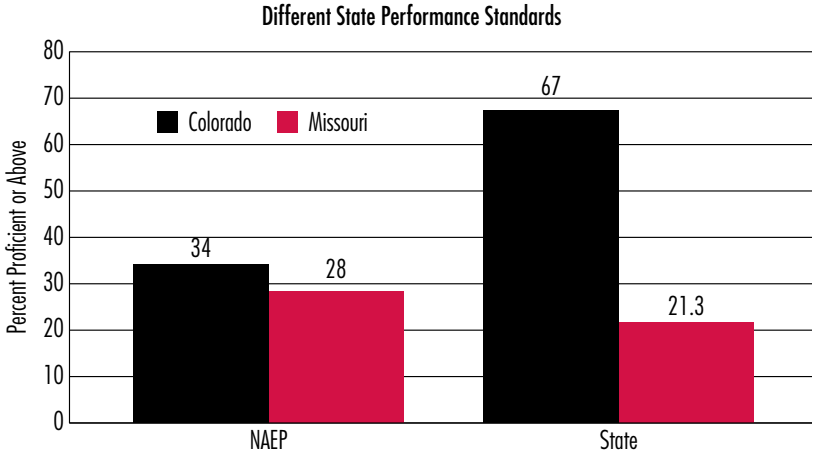
As an example of how measurement can go wrong, consider the development of assessments used to date under the NCLB system. When NCLB was introduced, each state was permitted to develop and use its own set of assessments for tracking student progress. What resulted was a set of 50 very different assessments.

To evaluate the differences across states, student performance on state assessments can be compared with student performance on the National Assessment of Educational Progress (NAEP). The NAEP is specifically designed to be consistent across states and over time. In fact, the test is considered the "nation's report card," a tool for evaluating what students in this country know. When researchers have compared how students in a state do on the NAEP relative to their performance on state assessments, the results are often striking (Fuller et al. 2006; Ho 2007).

As an example, consider Figure 1, taken from Linn (2005). The graph compares the percentage of students in Colorado and Missouri who score a level of "proficient" or above in 2003 on the NAEP and the states' own assessments. There are large differences in how students perform on various tests. While the number of students scoring at least "proficient" on the Missouri assessment is relatively similar to the number of students in Missouri who score "proficient" or above on the NAEP, students in Colorado are almost twice as likely to score at the "proficient" level or higher on the Colorado assessment than they are on the NAEP. Some have considered this evidence that some states set standards low in order to improve the proficiency rates of their students.

Concern about the differences in achievement measurement across states led to the design of the recently introduced Common Core State Standards, which have already been adopted by 30 states.[1] Relative to the existing system, this is a step in the right direction. But concerns exist about whether standardized tests measure all of the relevant skills. While we should certainly be interested in teaching children the math and reading skills that

**Figure 1.**
**Student Performance on State Assessments Relative to Performance on the National Assessment of Educational Progress**




1. *http://www.corestandards.org/in-the-states*

will help them be successful in life, and will in turn help our country's economy continue to grow, there are other skills that are important for these outcomes, too. For example, it has long been considered the purview of public schools to create an informed group of future voters and taxpayers. This is the rationale for social studies classes. Yet social studies is not a subject most schools are evaluating with their assessments.

Additionally, in school many children learn social and other skills, like patience, logic, and creative thinking, important for future life success but not measured adequately with current tests. There is a growing body of evidence that test scores do not tell the whole story when it comes to life success. Evidence from both the Perry Preschool Study and the Tennessee STAR class-size experiment shows that positive interventions early in life have little effect on children's scores on standardized tests many years later (in high school) (Cuhna and Heckman 2010; Chetty et al. 2010). Yet, the benefits of these interventions did translate into important improvements in longer-term life outcomes: increased wages, decreased crime rates, better health, etc. This evidence should caution us against placing too much weight on the standardized tests currently used and encourage us to use these scores in combination with other measures of teacher quality such as parental evaluations and in-class performance evaluations by trained assessors.

## Not All Grades and Subjects Are Subject to Testing

One of the downsides of using standardized tests to evaluate teachers is that most students take standardized tests only in some grades and some subjects. The NCLB Act of 2001 required that student achievement in reading and math be measured in grades three through eight. Many states have gone beyond this requirement to test students' knowledge in other subjects, such as science and writing, and at other grade levels, from pre-kindergarten to graduation. Yet many teachers still teach in subjects or grade levels that are not evaluated by standardized tests. As a result, evaluations of these teachers cannot include a measure of their own value-added score (or any other individual teacher-level standardized test measure).[2]

That some subjects or grade levels are not measured alters incentives of people entering teaching and of teachers already in the system. Teachers may feel as though standardized testing decreases their autonomy or otherwise makes the working conditions of a particular teaching position less desirable than a similar position in another grade or subject matter. If this is the case, teachers will find it less attractive to enter or stay in positions subject to standardized testing. On the other hand, if standardized testing offers a system for rewarding good teachers and is a part of a process of improving school functioning, teachers may find it more attractive to be in a classroom subject to assessment. Any change in incentives may be magnified in schools subject to increased pressures of accountability, e.g., those at risk of failing to meet requirements.

The early evidence on how standardized testing-based accountability systems affect teacher turnover and entry is mixed. One study found that teacher turnover is lower in assessed grades and subjects and that the entering teachers have more experience, a characteristic linked to teacher effectiveness (Boyd et al. 2008). Meanwhile, other researchers have found that turnover in low-performing schools is higher under accountability systems, potentially exacerbating the already acute problem of how to staff low-income and under-performing schools (Clotfelter et al. 2004; Feng et al. 2010). Since the studies occurred at different times and in different

2. Predicated on the idea that teachers can help one another learn to be better teachers, some districts and states have begun including a school-level measure of student achievement in the assessment of teachers in grades or subjects without testing. However, it is unlikely that a social studies teacher in grade six has much influence on her peers, especially those who teach in much higher or lower grades or those who teach quite different material, such as math.

states, their conflicting results suggest the particular parameters of standardized testing-based accountability regimes may have a lot to do with their effect on teacher turnover. Furthermore, despite their different results, taken together the studies imply accountability has changed teacher incentives in ways that will have effects on students. Parents, policymakers, and practitioners have to determine whether the benefits of accountability based on standardized testing are worth the potential side effects in subjects and grades without assessment and on students in low-performing schools.

## The Devil Is in the Details

Even when using tested methods like VAM, a lot of things can "go wrong" when doing statistical analysis of teacher effectiveness. The *Times* has suggested it will report the VAM measure for individual teachers on its website. I encourage more caution in doing so than was taken in the statistics reported in the newspaper's initial article. Figure 2 reproduces a graphic from the article comparing the performance of two teachers. Although it's not clear from the caption, one assumes from the description that the teachers' percentile ranks at the beginning and end of a particular year are being compared. If so, this is extremely problematic.

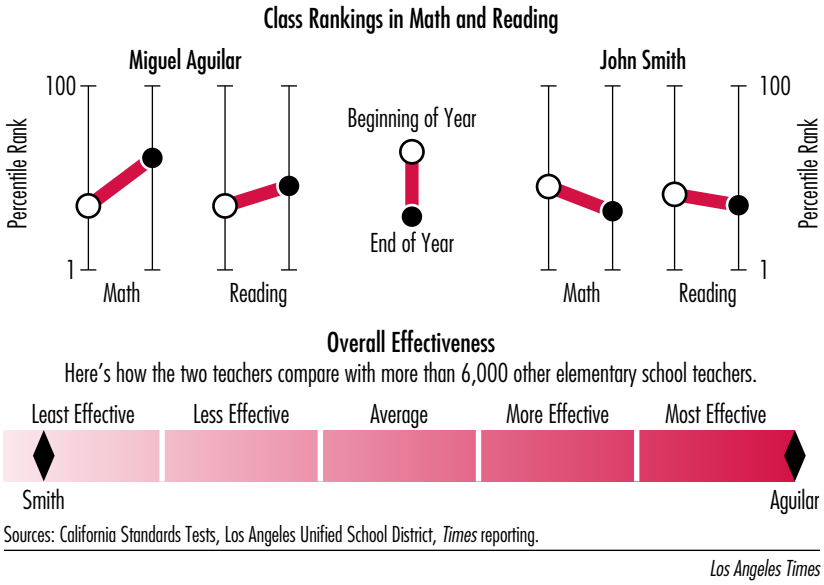
For one thing, the test scores, like most other statistics, are measured with error. Something as simple as a dog barking outside of the classroom on the day of the test could distract students and cause them to score lower on a test than they would have were the dog not there. If the dog had been sitting outside the classroom of John Smith, the decreases in the test scores of his students in the given year seen in Figure 2 would have nothing to do with his effectiveness. It would only imply that most other classrooms had not had such a distracting force outside their testing rooms.

Similarly, the size of a teacher's class is an important factor in the reliability of VAM. As an extreme example, consider a one-room schoolhouse with 20 children. If just a few of the children who enter the school in a given year are exceptionally bright, the average test score of the students in the schoolhouse will be very different than if those same entering students had learning needs. In an attempt to deal with this limitation, the *Times* limited its analyses to teachers with at least 60 students per year, but it's easy to imagine a few students throwing off the distribution of test scores of even 60 students. By comparing the test scores of a teacher's students with their own test scores a year earlier, the value-added models account for some of this natural cohort-level variation

*continued on flap...*

## Figure 2. Two Teachers

Two fifth-grade classes at Broadous Elementary School in Pacoima study the same lessons but end the year far apart. The difference is their teachers: Miguel Aguilar is one of the most effective in raising student test scores. John Smith is one of the least.




Sources: California Standards Tests, Los Angeles Unified School District, *Times* reporting.

*Los Angeles Times*

in classroom achievement. However, learning rates across students can also vary naturally in ways that affect value-added modeling. Consider the case if just a few of Mr. Smith's students were experiencing life events known to affect student achievement, like parental divorce or a recent move.

All of the above arguments about measurement point to one conclusion: reported scores of teacher effectiveness and the accountability systems based on them should be measured using multiple years of scores. Further, when reporting measures of teacher effectiveness, it is imperative to provide information about the quality of these measures. One way statisticians measure the reliability of an estimate, like the value-added measure, is to report confidence intervals. Looking at Figure 2, it's not hard to imagine confidence intervals on the reading scores of the two teachers that would suggest they are equally effective in the classroom.

## Conclusion

The LA Times has done the residents of the city of Los Angeles and the rest of the country a good service by drawing attention to the importance of individual teachers for children's education. Though the research on this matter has come a long way, we are still far from perfect solutions for how to measure teacher effectiveness and how to use those measurements to inform parents, policymakers, practitioners, and taxpayers. A good deal more careful thought should go into how these measurements are used. Such careful thought should be followed up by experimental implementation, in which methods of measurement and reporting are constantly evaluated and improved to suit the needs of students and taxpayers.

## References


Boyd, Donald J., Hamilton Lankford, Susanna Loeb, and James H. Wyckoff (2008). "The Impact of Assessment and Accountability on Teacher Recruitment and Retention: Are There Unintended Consequences?" *Public Finance Review* 36(1): 88–111.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Schanzenbach, and Danny Yagan (2010). "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence From Project STAR." Mimeo, Harvard University, *http://obs.rc.fas.harvard.edu/chetty/STAR_slides.pdf.*

Clotfelter, Charles T., Helen F. Ladd, Jacob L. Vigdor, and Roger Aliaga Diaz (2004). "Do School Accountability Systems Make It More Difficult for Low?Performing Schools to Attract and Retain High-Quality Teachers?" *Journal of Policy Analysis and Management* 23(2): 251–271.

Cuhna, Jesse, and James Heckman (2010). "Investing in Our Young People." National Bureau of Economic Research, Working Paper Number 16201.

Feltch, Jason, Jason Song, and Doug Smith (2010). "Who's teaching L.A.'s kids?" *The Los Angeles Times,* August 14, 2010.

Feng, Li, David Figlio, and Tim Sass (2010). *School Accountability and Teacher Mobility.* National Center for the Analysis of Longitudinal Data in Education Research, Working Paper Number 47.

Fuller, B., K. Gesicki, E. Kang, and J. Wright (2006). *Is the No Child Left Behind Act working? The reliability of how states track achievement.* University of California, Berkeley: Policy Analysis for California Education, Working Paper 06–1.

Ho, Andrew (2007). "Discrepancies Between Score Trends from NAEP and State Tests: A Scale-Invariant Perspective." *Educational Measurement: Issues and Practice* 26(4): 11-20.

Linn, Robert (2005). "Fixing the No Child Left Behind Accountability System." National Center for Research on Evaluation, Standards, and Student Testing, Policy Brief Number 8.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain (2005). "Teachers, Schools, and Academic Achievement." *Econometrica* 73(2): 417–458.

Rockoff, Jonah E. (2004). "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data." *American Economic Review* 94(2): 247–252.

# SIEPR

## About SIEPR

The Stanford Institute for Economic Policy Research (SIEPR) conducts research on important economic policy issues facing the United States and other countries. SIEPR's goal is to inform policymakers and to influence their decisions with long-term policy solutions.

## Policy Briefs

SIEPR policy briefs are meant to inform and summarize important research by SIEPR faculty. Selecting a different economic topic each month, SIEPR will bring you up-to-date information and analysis on the issues involved.

SIEPR policy briefs reflect the views of the author. SIEPR is a non-partisan institute and does not take a stand on any issue.

## For Additional Copies

Please see SIEPR website at *http://SIEPR.stanford.edu.*

# SIEPR *policy brief*

A publication of the
Stanford Institute for Economic Policy Research
Stanford University
366 Galvez Street
Stanford, CA 94305
MC 6015

Non-Profit Org.
U.S. Postage
**PAID**
Palo Alto, CA
Permit No. 28