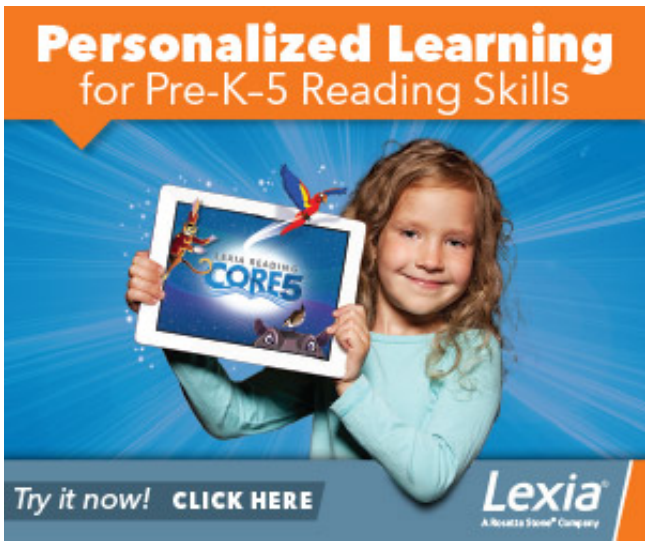
EDUCATION WEEK
Published Online: May 13, 2014
Published in Print: May 21, 2014, as **Researchers Advise Caution on Value-Added Models**

Studies Highlight Complexities of Using Value-Added Measures

By Holly Yettick

Back to Story



As a growing number of states begin evaluating teachers' effectiveness based on changes in their students' test scores, academic researchers are raising more questions about such "value-added" models.

In a study published Tuesday in the peer-reviewed journal *Educational Evaluation and Policy Analysis*, Morgan Polikoff, an assistant education professor at the University of Southern California in Los Angeles, and Andrew Porter, an education professor at the University of Pennsylvania in Philadelphia, **found no association** between value-added results and other widely accepted measures of teaching quality, such as the degree to which instruction is aligned with state standards or the contents of assessments. Nor did the study find associations between "multiple measure" ratings, which combine value-added measures with observations and other factors, and the amount and type of content covered in classrooms. That finding is potentially important because many states have responded to the Race to the Top grant competitions and other federal initiatives by adopting multiple-measure evaluation systems for teachers.

For their study, the researchers drew on a subset of data from 327 4th and 8th grade teachers in six school districts. The data were collected for the **Measures of Effective Teaching study**, which was funded by a $45 million grant from the in the Bill & Melinda Gates Foundation. The new study was supported by a $125,000 grant from the Gates Foundation.

Mr. Polikoff suggested that the study's findings could represent something like the worst-case scenario for the correlation between value-added scores and other measures of instructional quality.

"In some places, [value-added measures] and observational scores will be correlated, and in some places they won't," he wrote in an email. "These correlations will depend on things like the content and quality of the test, the type of [measures] used, the properties of the observation system. When the correlations are zero or near-zero, as we found, the questions are … 'What can we do about it?' and, 'How can we use this information to improve instruction and achievement?'"

Mr. Polikoff suggested that, despite the questions raised by his findings, value-added models, often referred to as VAMs, may still provide useful information for teacher evaluation and improvement

"I think that 'good teaching' probably has many different dimensions," he said. "The coverage of standards-aligned content. The ability to raise test scores. The ability to improve other noncognitive outcomes. ... These things may or may not be highly correlated."

"So I think that what [value-added models] are measuring is one narrow slice of what it means to be an effective teacher."

Many Facets

Indeed, a study published in February in the peer-refereed *American Educational Research Journal* found that **principal evaluations of teachers may capture results above and beyond** those that are assessed by value-added measures.

For that study, Douglas Harris, an economics professor at Tulane University in New Orleans, and his co-authors asked 30 Florida principals to rate teachers at their schools in 2005 and 2006 and then compared those ratings with value-added scores. They found that teachers with very good value-added ratings were more likely to get very good ratings from principals.

Overall, however, the principals' ratings and the value-added ratings were only weakly correlated. One reason is that both principal ratings and value-added models contain some level of measurement error, Mr. Harris said.

But Mr. Harris also offered another explanation.

"You can think of principal evaluations and value-added as measuring two different elements of 'quality instruction' in the same way that temperature and humidity are two key elements of quality weather," Mr. Harris said. "Sometimes temperature is high and humidity is low, sometimes these are reversed, and sometimes they are similar. So, it's no surprise that principal evaluations differ from value-added [ratings]."

For example, principals may give teachers an "A for effort" just as teachers do with students. They might be impressed by teachers who tried to improve their instruction by seeking professional development, and be concerned when burnout or personal conflicts appear to reduce the amount of time devoted to the classroom.

"It is easy to see why this might be," Mr. Harris said. "Anyone who runs an organization wants people to work hard and learn new things. On the other hand, some of the efforts principals prioritize don't seem to have much connection to student learning."

In addition to raising questions about the sometimes weak correlations between value-added assessments and other teacher-evaluation methods, researchers continue to assess how value-added models are created, interpreted, and used.

In a study that appears in the April issue of the *American Educational Research Journal*, Noelle A. Paufler and Audrey Amrein-Beardsley, a doctoral candidate and an associate professor, respectively, at Arizona State University in Tempe, respectively, **conclude that elementary school students are not randomly distributed into classrooms**. That finding is significant because random distribution of students is a technical assumption underlying some value-added models.

Even when value-added models do account for nonrandom classroom assignment, they typically fail to consider behavior, personality, and other factors that profoundly influenced the classroom-assignment decisions of the 378 Arizona principals surveyed. That omission, too, can bias value-added results.

Perhaps most provocative of all are the preliminary results of a study that uses value-added modeling to assess teacher effects on a trait they could not plausibly change, namely, their

students' heights. The results of that study, led by Marianne P. Bitler, an economics professor at the University of California, Irvine, have been presented at multiple academic conferences this year.

The authors found that teachers' one-year "effects" on student height were nearly as large as their effects on reading and math. While they found that the reading and math results were more consistent from one year to the next than the height outcomes, they advised caution on using value-added measures to quantify teachers' impact.

"Taken together, our results provide a cautionary tale for the interpretation and use of teacher VAM estimates in practice," Ms. Bitler and her colleagues wrote in a summary of **the study for the March conference of The Society for Research on Educational Effectiveness**, held in Washington.

"We find that—simply due to chance—teacher effects can appear large, even on outcomes they cannot plausibly affect," the authors wrote. "The implication is that many value-added studies likely overstate the extent to which teachers differ in their effectiveness, although further research is needed. Furthermore, users of [value-added measures] should take care to ensure their estimates reflect … teacher effectiveness and are not driven by noise."

Alternative Uses

Still another perspective on value-added models and their potential uses and misuses is provided by an article scheduled to appear in the June issue of the peer-refereed *Economics of Education Review*.

For the forthcoming study, a team of Brigham Young University researchers, led by assistant professor Scott Condie, drew on reading and math scores from more than 1.3 million students who were 4th and 5th graders in North Carolina schools between 1998 and 2004.

The authors found that between 15 percent and 25 percent of teachers were **misranked by typical value-added assessments** because the models did not consider that some teachers were more effective at teaching certain subjects (for example, reading vs. math) or certain types of students (higher vs. lower-achieving).

"[A] teacher who is fired may generate more social value than one who is retained," Mr. Condie and his co-authors write. "Thus, part of the critique of using value-added measures for personnel policies is undoubtedly correct; such policies will unfairly fire a large number of the wrong teachers."

A related challenge, the authors note, is that past research suggests that the pool of potential replacements is, on average, of lower quality than the pool of current teachers.

So the authors used the data to conduct two main types of simulations. First, they asked what might happen if North Carolina were to fire the "bottom" 10 percent of all teachers, as measured by value-added scores. As they expected, students' test scores rose.

However, they then asked another question: What would happen if, rather than using value-added data to fire the "bottom" teachers, the state instead used the data to match all teachers with the subjects and

RELATED BLOG

Visit this blog.

students for which they had demonstrated the strongest results? That approach, they found, also increased test scores—in fact, by much more than the teacher-firing model did.

"Our results suggest that employers might realize greater gains by increasing the specialization of their employees' tasks rather than attempting to replace them with hypothetically better employees," the Brigham Young authors conclude.

Of course, some districts and states already have policies that require them to use value-added models to help make high-stakes decisions about firing, tenure, and pay.

Asked what advice he would offer policymakers who have adopted or are considering adopting such measures, Mr. Polikoff of the University of Southern California replied: "I'd think about a way to design a system that takes each measure of effective teaching on its own terms, rather than a system that forces multiple, loosely-related measures into a single index for the purposes of making a decision,"

"Whatever system they choose, they should take it slow, thoughtfully study how this is all working (or not), and not be overly prescriptive," he added.

Vol. 33, Issue 32, Pages 10-11