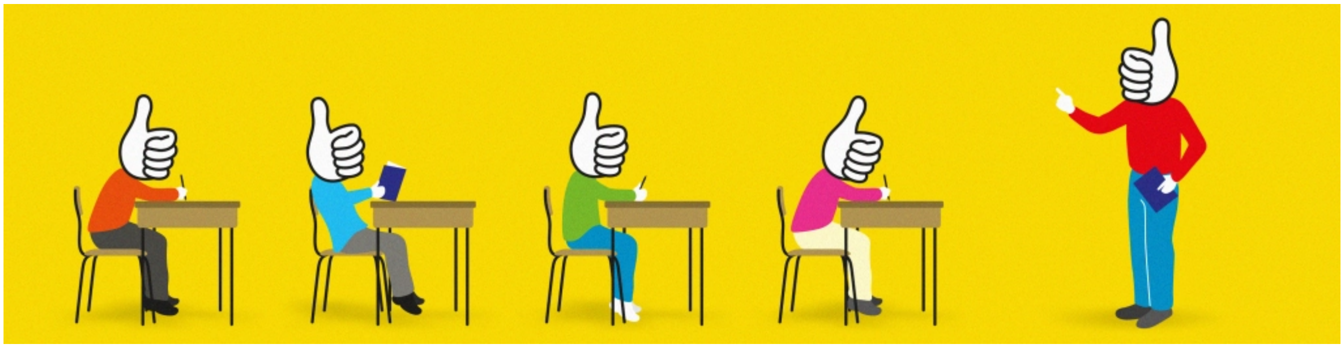

FINDINGS | 6:45 AM | JUL 20, 2015

The Science Of Grading Teachers Gets High Marks

By ANDREW FLOWERS

Is evaluating teachers an exact science? Many people — including many teachers and their unions — believe current methods are often too subjective and open to abuse and misinterpretation. But new tools for measuring teacher effectiveness have become more sophisticated in recent years, and several large-scale studies in New York, Los Angeles and North Carolina have given those tools more credibility. A new study released on Monday furthers their legitimacy; and as the science of grading teachers advances, it could push for further adoption of these tools.

This evolving science of teacher evaluation was recently thrust into public controversy when, in 2012, nine students sued the state of California, claiming its refusal to fire bad teachers was harming disadvantaged students. To claim that certain teachers were unambiguously bad, and that the state was responsible, the plaintiffs relied on relatively new measures of teacher effectiveness. In that case, *Vergara v. California*, several top-notch economists testified for each side as expert witnesses, arguing the merits of these complex statistics. In June 2014, the judge ruled that California's teacher-tenure protections were unconstitutional, a victory for the plaintiffs. Gov. Jerry Brown is appealing, and a similar case has begun in New York state.

But the economists on both sides of the *Vergara* case are still engaged in cordial debate. On one side is Raj Chetty of Harvard University, John Friedman of Brown University and Jonah Rockoff of Columbia University — hereafter referred to as "CFR" — who authored two influential papers published last year in the *American Economic Review*; Chetty testified for the plaintiffs in the case. On the other side is Jesse Rothstein, of the University of California at Berkeley, who published a critique of CFR's methods and supported the state in the Vergara case.

On Monday, to come full circle, the CFR researchers published a reply to Rothstein's criticisms.

At the center of this debate are evaluation models that try to isolate the educational value added by individual teachers, as measured by their students' standardized-test scores relative to what one would expect given those students' prior scores. The hard part, as Friedman says, is to "make sure that when you rate a teacher, that you actually rate what the teacher has done, and not whether they had a bunch of very poor or very rich students."

The CFR researchers — like the plaintiffs in the *Vergara* case — claim that these so-called "value added" models accurately isolate a teacher's impact on students, but Rothstein and critics say that value-added models, although improved, are still biased by factors outside the teacher's control.

In their pioneering papers published last year, the CFR researchers used massive data sets, covering millions of students over decades, to test whether teacher VA scores could accurately predict students' test scores. They argue they do, when done right, and thus can be used to winnow the good teachers from the bad. In CFR's method, teachers are judged from a baseline of the students' prior-year test scores, and by linking student scores to the tax records of parents, they are able to adjust for family characteristics like parents' income and the mother's age at childbirth. To many in the field, their approach is the cutting-edge science of teacher evaluation.

But Rothstein is still unconvinced. He claims that teacher value-added scores, even as tested by CFR, are biased. To understand why, step back and imagine the ideal scientific test: a randomized experiment. In order to perfectly isolate the effect of a teacher on a student's test scores — setting aside whether higher test scores is the right goal –– students would need to be assigned to teachers randomly.

Many people may think that students are already randomly assigned to teachers, but that's not always the case. Researchers on both sides of this debate have found some sorting *within* schools when they examine the data closely. It might be that wealthier, better-educated parents are effective at lobbying principals to place their kid in a better teacher's classroom. Because of this sorting, extra precaution — such as controlling for the students' demographic backgrounds — is needed.

Several experiments that randomly assign students have been done, but the studies lacked large enough sample sizes to precisely prove whether the resulting VA metrics were unbiased. Large, fully randomized experiments can be impractical in education research.

To approximate random assignment, the CFR researchers looked at a common school situation: teacher switching. This turned out to be the researchers' real innovation**.** Because teachers sometimes switch classrooms or schools, the researchers were able to use this switching to test whether the value-added measures accurately predict student performance. While short of being a true random experiment, the researchers say this quasi-experimental approach examining instances of teacher switching (with proper adjustments) can be as good as random.

For example, let's say Mrs. Smith teaches 6th grade math, and according to the value-added scores she's a 90th percentile teacher. But as recorded in CFR's massive database, Mrs. Smith leaves her school the following year, and the next group of 6th graders are taught by Mr. Johnson, who is merely a 50th percentile teacher. After controlling for how this new crop of students did on their 5th grade math exams, and holding constant their demographic characteristics, Chetty, Friedman and Rockoff can ask: How did Mr. Johnson's 6th grade class score on their end-of-year math exam relative to Mrs. Smith's class the year before? If the value-added measures accurately predict student scores, they would be considered unbiased, and according to CFR's analysis, they do and are.

As they put it, "VA [value added] accurately captures teachers' impacts on students' academic achievement." The implication being, school administrators can legitimately use value-added scores to hire, fire and otherwise evaluate teacher performance. Unsurprisingly, such control is fiercely resisted by teacher's unions. And they can point to Rothstein's work, which is critical of these metrics, to support their case.

Rothstein's overarching critique of CFR is that their quasi-experimental research design — where teacher-switching is used to test whether the value-added metrics are biased — is not actually a good approximation of a randomized experiment. In Rothstein's reply to CFR's original research, he claims that teacher switching is correlated with students' prior-year test scores. This could be the case when, for example, an exiting teacher is replaced by one with a higher value-added score, but simultaneously the next crop of students are also better prepared. This would bias the teacher value-added scores. What seems like a better teacher improving their students' scores isn't really — the students had a better baseline level of preparation going in.

But in their newly published response, the CFR researchers point out that Rothstein might just be picking up on a statistical artifact, not a real problem. Rothstein acknowledges this response "could be true," but maintains that he is unconvinced the CFR researchers have really proven their method is free of bias.

To get more perspective, I reached out to Thomas Kane, of Harvard University, and Douglas Staiger, of Dartmouth College — two economists with years of experience in this field. They were authors, with a third researcher, of a replication of CFR's method using data from Los Angeles public schools. In this debate, they view CFR's response as more convincing than Rothstein's criticism. After all, using CFR's methods on LA data, they got nearly the same result.

Kane, like Chetty, has been an expert witness in supporting value-added models; and his public writings confirm these leanings. Staiger, though, is one of the few researchers I spoke with who didn't have a conflict of interest (having not been an expert witness). "I'm convinced by the CFR response," Staiger said, and he doesn't believe their research design is flawed, as Rothstein claims.

And there is a broader point in favor of CFR's work: Their numbers are being replicated in many different settings. Even in Rothstein's paper critiquing their method, he replicated their results using data from North Carolina public schools. "I'm not aware of another area of social science where there has been so much replication, in such a short time, and they've all found the same result," Kane said. On the consistency of replicability, Staiger said "it's just astounding, actually." Even Rothstein grants this: "Replication is an extremely important part of the research process … I think this is a great success, that these very complex analyses are producing similar results."

"It's almost like we're doing real, hard science here," Friedman said. Well, almost. But by the standards of empirical social science — with all its limitations in experimental design, imperfect data, and the hard-to-capture behavior of individuals — it's still impressive. The honest, respectful back-and-forth of dueling empirical approaches doesn't mean the contentious nature of teacher evaluation will go away. But for what has been called the "credibility revolution" in empirical economics, it's a win.

Andrew Flowers is FiveThirtyEight's quantitative editor. | ✉ | *@andrewflowers*

FILED UNDER EDUCATION, FINDINGS, SCHOOLS, TEACHER EVALUATIONS