Carnegie Foundation for the Advancement of Teaching

DO DIFFERENT VALUE-ADDED MODELS TELL US THE SAME THINGS?

DAN GOLDHABER
AND **RODDY THEOBALD**
CENTER FOR EDUCATION DATA & RESEARCH
UNIVERSITY OF WASHINGTON BOTHELL

CARNEGIE KNOWLEDGE NETWORK
What We Know Series:
Value-Added Methods and Applications

ATIL ADVANCING • TEACHING • IMPROVING • LEARNING

HIGHLIGHTS:

- Statistical models that evaluate teachers based on growth in student achievement differ in how they account for student backgrounds, school, and classroom resources. They also differ by whether they compare teachers across a district (or state) or just within schools.

- Statistical models that do not account for student background factors produce estimates of teacher quality that are highly correlated with estimates from value-added models that *do* control for student backgrounds, as long as each includes a measure of prior student achievement.

- Even when correlations between models are high, different models will categorize many teachers differently.

- Teachers of advantaged students benefit from models that do not control for student background factors, while teachers of disadvantaged students benefit from models that do.

- The type of teacher comparisons, whether within or between schools, generally has a larger effect on teacher rankings than statistical adjustments for differences in student backgrounds across classrooms.

INTRODUCTION

There are good reasons for re-thinking teacher evaluation. As we know, evaluation systems in most school districts appear to be far from rigorous.[1] A recent study [2] showed that more than 99 percent of teachers in a number of districts were rated "satisfactory,"[3] which does not comport with empirical evidence that teachers differ substantially from each other in terms of their effectiveness.[4,5] Likewise, the ratings do not reflect the assessment of the teacher workforce by administrators, other teachers, or students.[6]

Evaluation systems that fail to recognize the true differences that we know exist among teachers greatly hamper the ability of school leaders and policymakers to make informed decisions about such matters as which teachers to hire, what teachers to help, which teachers to promote, and which teachers to dismiss. Thus it is encouraging that policymakers are developing more rigorous evaluation systems, many of which are partly based on student test scores.

Yet while the idea of using student test scores for teacher evaluations may be conceptually appealing, there is no universally accepted methodology for translating student growth into a measure of teacher performance. In this brief, we review what is known about how measures that use student growth align with one another, and what that agreement or disagreement might mean for policy.

WHAT DO WE KNOW ABOUT THIS ISSUE?

There is a growing body of research that compares estimates of teacher quality produced by different models. These models, which consider student growth on standardized tests, fall roughly into four categories: "value-added models" that do not control for student background; models that do control

for student background; models that compare teachers within rather than across schools; and student growth percentile (SGP) models, which measure the achievement of individual students compared to other students with similar test score histories.[7]

Multiple modeling options for estimating teacher quality

School districts and states that want to use student test scores to inform teacher evaluations have a number of options. There are five large vendors that use varied approaches to translating student growth information into measures of teacher quality (see Table 1).[8] The methods used by three of them—the Value Added Research Center (VARC) at the University of Wisconsin, the American Institutes for Research (AIR), and Mathematica Policy Research—are difficult to summarize because each vendor tailors the approach for each client. But, generally, they all use value-added models.[9]

Table 1: Large Vendors that Estimate Teacher Effectiveness Using Student Test Scores

Vendor	Name of Model	Brief Description
American Institutes for Research (AIR)	Varied	In most situations, models control for student background
Mathematica	Varied	In most situations, models control for student background
National Center for the Improvement of Educational Assessment (NCIEA)	Student Growth Percentile (SGP) Models	Models a descriptive measure of student growth within a teacher's classroom
SAS	EVAAS	Models control for prior test scores but not other student background variables
Value Added Research Center (VARC)	Varied	In most situations, models control for student background

Value-added models are statistical models that generally try to isolate the contributions to student test scores by individual teachers or schools from factors outside the school's or teacher's control. Such factors may include prior test scores, poverty, and race. For instance, students in poorly financed schools, whose parents are not engaged in their education, often do poorly on tests; the value-added model controls for these sorts of factors.[10] There is debate[11] over whether value-added models accurately capture the true contributions (in statistical parlance, "causal estimates") of schools and teachers as opposed to simply identifying correlational relationships.[12]

The approaches of the other two primary vendors—SAS and the National Center for the Improvement of Educational Assessment (NCIEA)—are easier to define, since each specializes in a specific model. SAS uses the Education Value Added Assessment System (EVAAS).[13] The value-added models discussed in the academic literature tend to include controls for the socio-economic and demographic characteristics of students in order to account for achievement gaps between certain groups. EVAAS, by contrast, intentionally omits these controls. One justification for the omission is that including demographic characteristics differentiates expectations for students in certain groups. (We return to this important distinction below in the section on *What can't be resolved by empirical evidence on this issue*?)

NCIEA takes a different approach with the Student Growth Percentile (SGP) model,[14] often called the "Colorado Growth Model." Both the SGP model and value-added models use a statistical technique called regression to analyze the test score histories of students. Where value-added models purport to separate the contributions of teachers from other variables, the SGP model provides a student growth percentile for each student that shows their growth relative to other student with similar test-score histories. Therefore the SGP model is focused at the student level, and individual student scores are aggregated to the classroom level to obtain a measure of teacher performance; typically the median (or mean) student growth percentile for each teacher's students becomes the teacher's SGP score (the "average" growth of the teacher's students). Outputs from both types of models are currently being used as part of teacher evaluations systems. By design, SGP models do not purport to provide causal estimates of teacher effectiveness (though this does not necessarily imply that they are less accurate measures); they are intended as a descriptive measure of *what is* – of test score gains relative to other students who scored similarly in the past."[15]

Different models' approaches to isolating the effect of the teacher

Given the many options, school districts and states likely have a number of questions about which model is "right" for them. Which model, if any, provides a fair estimate of a teacher's contribution to student learning? In fact, do value-added estimates provide causal estimates at all? These are heated questions in the field, and they are unlikely to be resolved without more data. But different value-added models make different assumptions about how much variation in test scores should be attributed to teachers. The SAS EVAAS model removes variability due to students' previous test scores. The three other models also control for previous scores, but they often control for other factors, as well: Mathematica's DC IMPACT model removes variability due to differences in student background,[16] Mathematica's Pittsburgh model removes variability due to average classroom characteristics[17], and AIR's Florida VAM removes variability that may be due to school-wide factors. Again, NCIEA's SGP model, instead of removing variability statistically, explicitly controls for student background by comparing only students with similar test score histories. The extent to which these different modeling decisions matter depends on three factors: the correlation between each of these variables and student achievement; how inequitably students are distributed across different classrooms and schools; and how inequitably teacher quality is distributed within and across schools.[18] We return to this discussion below in the section on *What can't be resolved by empirical evidence on this issue?*

Correlations between value-added estimates from different models

A question that can be answered with empirical data is the extent to which estimates from selected models correlate with each other.[19] . Many studies[20] have calculated high correlations (mostly greater than 0.9) between estimates from models that control only for prior student test scores (such as SAS EVAAS),[21] control for student background (such as DC's IMPACT), and control for average classroom characteristics (such as Pittsburgh's system). Importantly, two studies[22] calculate higher correlations between estimates that use different models than between estimates that use different exams.

It is only recently that researchers have begun to compare estimates generated by traditional value-added and SGP models. Wright (2010)[23] compares SGP estimates to estimates produced by the SAS EVAAS approach, while Goldhaber et al (2012)[24] compare SGP estimates to estimates from the full range of value-added models discussed above. Both studies find what might be seen as surprisingly

high correlations (around 0.9) between estimates from value-added and SGP models. [25] We say "surprisingly" both because the two types of models have different motivations and because the kinds of student background variables that are in value-added models are often found to influence student achievement.[26]

Most model options result in estimates that correlate highly with one another; however, there is a critical decision that results in estimates with far lower correlations – how teachers should be compared to each other. The two most common choices are to compare teachers across all schools in a district, or to compare teachers only to other teachers in the same school. A few studies[27] compare estimates from these two types of models and find correlations closer to 0.5.

There are two potential explanations for why within-school comparisons change estimates of teacher effectiveness so much. First, schools themselves may make significant contributions to student learning that get attributed to teachers when we don't account for school factors. Alternatively, teacher quality may be inequitably distributed across schools, meaning that below-average teachers with a lot of below-average peers look a lot better when comparisons are made within schools, while above-average teachers with a lot of above-average peers look worse.[28] Given the trade-off between these two factors, it is difficult to know which type of model is "better."[29] We will address this question more fully below in *What can't be resolved by empirical evidence on this issue?*[30]

The impact of model choice on teachers' effectiveness ratings

While the correlations tell us the degree of agreement of effectiveness estimates, they do not provide the kind of contextual information that individual teachers likely care about. Specifically, teachers want to know how they would rank under different modeling approaches and in what effectiveness category they would fall.[31]

We illustrate the relationship between model correlation and teacher classification in Table 2.[32] In particular, we place teachers into performance quintiles[33] based on how they would rank under different models and compare that rating to the ratings that would result for the same teachers under different models.[34] Panels A-C represent math performance and Panels D-F represent reading performance. Each panel compares the rating of teachers using a value-added model with prior test scores and student covariates to placements from another model: (1) a value-added model that includes only a prior test score (Panels A and D); (2) the SGP model (Panels B and E); and (3) a value-added model that makes within-school comparisons (Panels C and F).[35]

Table 2: Transition Matrices and Correlations for Different Effectiveness Estimates

Panel A. Math	Correlation = 0.97	VAM with prior test score				
		Q1 (Lowest)	Q2	Q3	Q4	Q5 (Highest)
VAM with prior test score and student covariates	Q1 (Lowest)	17.2%	2.7%	0.0%	0.0%	0.0%
	Q2	2.7%	13.7%	3.6%	0.1%	0.0%
	Q3	0.1%	3.4%	12.9%	3.5%	0.0%
	Q4	0.0%	0.2%	3.4%	13.8%	2.6%
	Q5 (Highest)	0.0%	0.0%	0.1%	2.6%	17.3%
Panel B. Math	**Correlation = 0.91**	**Student Growth Percentiles**				
		Q1 (Lowest)	Q2	Q3	Q4	Q5 (Highest)
VAM with prior test score and student covariates	Q1 (Lowest)	15.4%	4.2%	0.4%	0.0%	0.0%
	Q2	4.1%	10.2%	5.0%	0.6%	0.0%
	Q3	0.6%	5.0%	9.4%	4.7%	0.4%
	Q4	0.0%	0.9%	5.3%	10.2%	3.6%
	Q5 (Highest)	0.0%	0.0%	0.5%	4.2%	15.3%
Panel C. Math	**Correlation = 0.55**	**VAM with within-school comparison**				
		Q1 (Lowest)	Q2	Q3	Q4	Q5 (Highest)
VAM with prior test score and student covariates	Q1 (Lowest)	9.0%	5.6%	3.1%	1.6%	0.8%
	Q2	5.0%	5.4%	4.6%	3.3%	1.7%
	Q3	3.1%	4.4%	5.0%	4.5%	3.0%
	Q4	1.9%	3.0%	4.5%	5.5%	5.0%
	Q5 (Highest)	0.9%	1.5%	2.7%	5.2%	9.4%
Panel D. Reading	**Correlation = 0.91**	**VAM with prior test score**				
		Q1 (Lowest)	Q2	Q3	Q4	Q5 (Highest)
VAM with prior test score and student covariates	Q1 (Lowest)	15.4%	4.0%	0.5%	0.1%	0.0%
	Q2	4.0%	10.4%	4.7%	0.9%	0.1%
	Q3	0.5%	4.7%	9.6%	4.6%	0.6%
	Q4	0.1%	0.8%	4.7%	10.5%	3.9%
	Q5 (Highest)	0.0%	0.1%	0.5%	4.0%	15.4%
Panel E. Reading	**Correlation = 0.81**	**Student Growth Percentiles**				
		Q1 (Lowest)	Q2	Q3	Q4	Q5 (Highest)
VAM with prior test score and student covariates	Q1 (Lowest)	13.8%	4.5%	1.4%	0.2%	0.0%
	Q2	5.5%	7.3%	5.3%	1.7%	0.2%
	Q3	1.8%	5.0%	7.2%	4.6%	1.4%
	Q4	0.4%	2.0%	5.7%	7.1%	4.8%
	Q5 (Highest)	0.0%	0.4%	1.8%	4.8%	12.9%
Panel F. Reading	**Correlation = 0.52**	**VAM with within-school comparison**				
		Q1 (Lowest)	Q2	Q3	Q4	Q5 (Highest)
VAM with prior test score and student covariates	Q1 (Lowest)	8.8%	5.5%	3.0%	1.7%	1.2%
	Q2	4.9%	5.5%	4.6%	3.1%	2.0%
	Q3	3.2%	4.4%	4.9%	4.4%	3.2%
	Q4	2.0%	3.0%	4.5%	5.4%	4.8%

	Q5 (Highest)	1.2%	1.6%	3.0%	5.4%	8.8%

If there were complete agreement between two different approaches to estimating teacher performance, we would expect each of the shaded boxes along the diagonal to contain 20 percent of the teachers, since both models would perfectly agree on how to rate each teacher. Clearly this is not the case for any of the comparisons. And, importantly, even models that are very strongly correlated—such as math value-added models with and without student covariates in (Panel B, r = 0.97)—show considerable movement between quintiles. For example, of the teachers identified in the bottom quintile by the value-added model with student covariates, over 13 percent move out of the bottom quintile when we control for student covariates. The movement becomes more pronounced as the correlations decrease, both as we compare the value-added model with student covariates to SGP models and within-school models, and as we make the same comparisons in reading. Strikingly, about 6 percent of teachers who are placed in the top quintile in reading by the value-added model with student covariates are placed in the bottom quintile by the value-added model that makes within-school comparisons, and vice versa.

The differential impact of model choice on teachers' effectiveness

Correlations also do not distinguish among the *types* of teachers affected by different types of models. Of particular concern is whether one model or another unduly affects teachers who work primarily with disadvantaged students. Table 3[36] compares the average percentile rankings of teachers in the most advantaged classrooms to the average percentile rankings of teachers in the least advantaged classrooms for different estimates of teacher effectiveness.

Table 3 demonstrates that SGP and value-added models that do not control for student covariates systematically favor teachers in advantaged classrooms. In reading, for example, the average percentile ranking for teachers in advantaged classrooms is 58.2 compared to 43.6 for teachers in disadvantaged classrooms when teacher effectiveness is calculated with a value-added model that does control for student covariates. But the gap is markedly wider for SGP models and value-added models that do not control for student covariates beyond previous test scores; it is 66.6 compared to 33.8 for SGP estimates, and 71.8 compared to 29.0 for value-added estimates that control only for prior student achievement.[37]

Table 3: Average Percentile Rankings in Advantaged and Disadvantaged Classrooms

Panel 1: Math	**Advantaged**	**Disadvantaged**
Student Growth Percentiles	60.7	41.1
VAM with prior test score	65.1	38.2
VAM with prior test score and student covariates	57.8	47.7
VAM with prior test score, student, and classroom covariates	60.1	46.6
VAM with within-school comparison	51.9	48.7
Panel 2: Reading	**Advantaged**	**Disadvantaged**
Student Growth Percentiles	66.6	33.8
VAM with prior test score	71.8	29.0
VAM with prior test score and student covariates	58.2	43.6
VAM with prior test score, student, and classroom covariates	60.3	42.8
VAM with within-school comparison	51.0	49.4

WHAT MORE NEEDS TO BE KNOWN ON THIS ISSUE?

We have discussed what is already known about how estimates from these models agree with each other and which teachers are affected by any differences that exist. There is an emerging consensus over the answer to the first question; the answer to the second is more preliminary. But in both cases, findings will have to be validated across different states and contexts.

There is another question that could be answered with the right empirical data: what evaluation systems that use student test scores actually lead to greater changes in student performance? Districts and states are presumably adopting new evaluation policies because they believe these policies could lead to better student achievement. Different places have adopted markedly different models, each with its own consequences and rewards. While much of the discussion has focused on teacher bonuses and dismissals, many districts are considering other uses of value-added models, including tying evaluation scores to professional development.[38] Once a few years have passed, researchers will be able to determine whether any of these systems have led to changes in the teacher workforce or in student achievement.

WHAT *CAN'T* BE RESOLVED BY EMPIRICAL EVIDENCE ON THIS ISSUE?

There are three important questions about the use of student growth measures that cannot easily be answered with existing empirical evidence:

1. Which model produces estimates that are the "fairest" to individual teachers?
2. What is the appropriate balance between accuracy and transparency for evaluation systems that use student test data?
3. Is it more appropriate to compare teachers within schools or across schools?

But, the three comparisons in Table 2 provide case studies for each of these questions.

The question of "fairness" is likely to be at the heart of any debate about teacher evaluation, and the debate over controlling for student characteristics other than prior test scores shows why it is hard to know which model is more "fair" to individual teachers. As we have seen, the evidence demonstrates that teachers who teach in advantaged classrooms benefit from models that do not control for student covariates like race and poverty, but policymakers may be reluctant to employ models that do control for these factors. This is because, given what we know about the relationship between these variables and student achievement, the model would expect low-income students to show lesser gains than high-income students. So a teacher of disadvantaged students could get a higher value-added measure than a teacher of advantaged students even though her class showed less actual growth. On the other hand, this sort of outcome may seem perfectly fair given that some teachers face greater obstacles than others given the readiness to learn of the students in their classrooms. Moreover, many administrators are understandably loath to use an evaluation system that may discourage teachers from working in disadvantaged classrooms.[39] It is purely a judgment call about which result

is fairer to individual teachers and how it may affect achievement goals. What is fair to teachers may not be fair to students, and vice versa.[40]

Differences between value-added and SGP rankings (Panels B and E of Table 2) illustrate the potential tradeoff between transparency and accuracy in an evaluation system. SGPs may not be designed to give causal estimates of teacher effectiveness, but they are understandably appealing to many policy makers and administrators because of the transparency of the resulting scores. It is much easier to explain to parents, for example, that a teacher's score is the median estimate of student growth in her class as opposed to a coefficient from a linear regression.[41] Again, we can argue that value-added estimates may be better designed to give causal estimates of teacher quality, but whether the transparency of SGPs outweighs this benefit is another matter for policy makers and administrators to judge.

Finally, the issue of whether to compare teachers within or across schools (see Panels C and F of Table 2) is another situation for which empirical evidence provides little guidance. With just a single year of data[42] there is no way to know whether differences in student performance across schools are due to school factors or differences in the average quality of teachers. A model that compares teachers to the average teacher across all schools produces estimates of teacher effectiveness that are combinations of teacher and school effects on student achievement.[43] But a model that compares teachers to the average teacher within a school assumes that teacher quality is distributed evenly across schools. It may also lead to competition rather than cooperation between teachers. Given that statistical models simply cannot distinguish between teacher and school effects with just one year of data, policymakers and administrators again must decide which model is most appropriate.

PRACTICAL IMPLICATIONS

How does this issue impact district decision making?

Our reading of the research shows that modeling choices do have an important impact on teacher rankings.[44] Of models that control for prior achievement, the high correlations between those that do and do not explicitly account for student characteristics might suggest that debates over covariate adjustments are misplaced. But high correlations can, as we show above, mask pretty large differences in the rankings of teachers in different classrooms. We would argue that the differences are meaningful enough for policymakers to be concerned when deciding what model to adopt.

The evidence about the impact of model specification on performance rankings raises another issue for consideration. Most districts and states have only just begun to use student growth measures to inform high-stakes decisions about teachers. It is likely that their methods for doing so will change. New York State, for example, plans to report SGP scores to teachers in 2011-12 and value-added scores beginning in 2012-13 in certain subjects and grades.[45] The empirical evidence suggests that, depending on the model adopted, this change could have consequences for individual teachers. So even though each of these evaluation systems may be superior to those now used elsewhere, the potential shifts in teacher rankings could serve to undermine the usefulness of both.[46]

The data needed to estimate teacher performance based on student growth are now widely available. This means that administrators who use these measures for high-stakes purposes could be confronted by teachers who could rightly argue, and point to empirical evidence, that their ranking would have been different under different assumptions. One cannot escape the fact that different models lead to different results, but the issue of *how* they do could be made clear to everyone when the models are being adopted. Transparency would encourage buy-in at the start and help prevent surprises later.[47]

ENDNOTES

[1] For a more comprehensive discussion, see:
Steven Glazerman, Dan Goldhaber, Susanna Loeb, Stephen Raudenbush, Douglas Staiger, and Grover J. "Russ" Whitehurst, "Passing Muster: Evaluating Teacher Evaluation Systems," (Brookings Institution, Washington, DC, 2011).
Dan Goldhaber, "When the Stakes are High, Can We Rely on Value Added?" Center for American Progress, December 2010.
Thomas Toch and Robert Rothman, "Rush to judgment: Teacher evaluation in public education," (Report for Education Sector, Washington, D.C., 2008).

[2] Daniel Weisberg, Susan Sexton, Jennifer Mulhern, and David Keeling, "The widget effect: Our national failure to acknowledge and act of differences in teacher effectiveness," (Report for the New Teacher Project, 2009.)

[3] There was slightly more spread in evaluations in districts using a broader range of ratings, but it was still about 95 percent of teachers who received one of the top two ratings in these districts.

[4] Daniel Aaronson & Lisa Barrow & William Sander, "Teachers and Student Achievement in the Chicago Public High Schools," Journal of Labor Economics, University of Chicago Press, 25, (2008): 95-135.
Dan Goldhaber and Michael Hansen, "Is it just a bad class? Assessing the stability of measured teacher performance," (Paper for Center for Education Data and Research, 2010).
Eric A. Hanushek, "The trade-off between child quantity and quality," Journal of Political Economy, 100, no.1, (1992): 84-117.
Steven G. Rivkin, Eric A. Hanushek, and John F. Kain, "Teachers, schools, and academic achievement," Econometrica, 73, no.2, (2005):417-458.

[5] Raj Chetty, John F. Friedman, and Jonah E. Rockoff, "Long-Term Impact of Teachers: Teacher Value-Added and Student Outcomes in Adulthood," National Bureau of Economic Research, (Working Paper #w17699, 2011).

[6] Brian Jacob and Lars Lefgren, "Principals as agents: Subjective performance assessment in education," Journal of Labor Economics, 26, no.1, (2007): 101-136. Also see Tucker (1997) and Weisburg et al (2009).
Bill and Melinda Gates Foundation, Working with teachers to develop fair and reliable measures of effective teaching, 2010.

[7] Student growth percentiles are aggregated to the classroom level as an assessment of teachers. Some models also account for school or classroom resources, such as class size, and/or measure classroom level characteristics, such as the average poverty level of the classroom.

[8] We use the term "teacher quality" to mean the estimated *ability of teachers to contribute in measurable ways to student gains on standardized tests*, and we treat this as synonymous with the terms "teacher performance" and "teacher effectiveness."

[9] Value-added models have long been used by researchers to assess the effects of schooling attributes (class size, teacher credentials, etc.) on student achievement. See, for instance,
Eric A. Hanushek, "The economics of schooling - production and efficiency in public-schools," *Journal of Economic Literature,* 24, no.3, (1986):1141-1177.

[10] Dale Ballou, William Sanders, and Paul Wright , "Controlling for student background in value-added assessment of teachers," Journal of Education and Behavioral Statistics, 29, no.1, (2004): 37-65.
Robert Gordon, Thomas J. Kane, and Douglas O. Staiger, *Identifying effective teachers using performance on the job*, (Report for Brookings Institution, Washington, D.C., 2006).

Daniel F. McCaffrey, J. R. Lockwood, Daniel Koretz, Thomas A. Louis, Laura S. Hamilton, "Models for value-added modeling of teacher effects, "Journal of Educational and Behavioral Statistics, 29, no.1, (2004): 67-101.

[11] See Chetty et al (2011) and:
Dan Goldhaber and Duncan Chaplin, "Assessing the "Rothstein Falsification Test." Does it Really Show Teacher Value-added Models are Biased?" (CEDR Working Paper, University of Washington, Seattle, WA, 2012-13).
Jesse Rothstein, "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement," Quarterly Journal of Economics 125, no.1, (2010): 175-214.

[12] For example, we might see that high-achieving students tend to be taught by teachers with Master's degrees. But the simple correlation between achievement and level of a teacher's degree would not tell us whether having a teacher with a Master's degree actually *leads* to higher achievement (whether it has a causal effect) or whether schools with high achievers just tend to hire more teachers with Master's degrees. Regardless, most value-added estimates are *designed* to identify causal relationships – to capture true teacher contributions to student learning. To read more about what is known on the influence of student characteristics on value added, see: McCaffrey, Daniel. Carnegie Knowledge Network, "Do Value-Added Methods Level the Playing Field for Teachers?" Last modified October 2012. URL = <http://carnegieknowledgenetwork.org/briefs/value-added/level-playing-field./>

[13] Paul Wright, John T. White, William L. Sanders, and June C. Rivers, SAS EVAAS statistical models, (SAS White Paper, 2010).

[14] Damian W. Betebenner, "Norm- and criterion-referenced student growth," Educational Measurement: Issues and Practice, 28, no. 4, (2009):42-51.

[15] Betebenner (2009)

[16] Beginning in the 2012-13 school year, the DC Impact model will also control for aggregated classroom characteristics (Eric Isenberg, personal communication, September 2012.)

[17] For more information about classroom peer effects, see:
Mary A. Burke and Tim R. Sass, "Classroom Peer Effects and Student Achievement," Journal of Labor Economics, 31, no. 1, (2013): 51-82.

[18] The best evidence we have on this question is from Kane and Staiger (2008), who use experimental data to argue that models that control for average classroom characteristics (like Mathematica's model for Pittsburgh Public Schools) provide the best accuracy. The Measures of Effective Teaching project will provide further evidence on this topic.

[19] Correlation can be measured with a Pearson correlation, which is a measure of linear agreement, or a Spearman correlation, which is a measure of how well the rankings of the two measures agree. The papers we review are inconsistent as to the measure of correlation they use, as well as whether they use shrunken or non-shrunken teacher effects. The results are robust to the measures used, however, so we do not discuss these distinctions further.

[20] See Ballou et al. (2004) and:
Douglas N. Harris, and Tim R. Sass, "Value-added models and the measurement of teacher quality," (Working paper, 2006.)
J. R. Lockwood, Daniel F. McCaffrey, Laura S. Hamilton, Brian M. Stecher, Vi-Nhuan Le and Jose Felipe Martinez, "The sensitivity of value-added teacher effect estimates to different mathematics achievement measures," Journal of Educational Measurement, 44, no.1, (2007): 47-67.
John P. Papay, "Different tests, different answers: The stability of teacher value-added estimates across outcome measures," American Educational Research Journal, 48, no.1, (2011): 163-193. doi: 10.3102/0002831210362589
Dan Goldhaber, Joe Walch, and Brian Gabele, "Does the model matter? Exploring the relationship between different student achievement-based teacher assessments," Statistics, Politics, and Policy, in press.

Mark Ehlert, Cory Koedel, Eric Parsons, and Michael Podgursky, "Selecting Growth Measures for School and Teacher Evaluations," (University of Missouri Working Paper 12-10, 2012).

[21] Often for multiple years of prior student tests, e.g. an assessment of student achievement in the 5th grade controlling for achievement in the 4th and 3rd grades.

[22] Lockwood et al (2007) and Papay (2011)

[23] S. Paul Wright, "An investigation of two nonparametric regression models for value-added assessment in education," (SAS White Paper, 2010).

[24] Dan Goldhaber, Joe Walch, and Brian Gabele, "Does the model matter? Exploring the relationship between different student achievement-based teacher assessments," Statistics, Politics, and Policy, in press.

[25] Ehlert et al (2012) find slightly smaller correlations for estimates of school effectiveness (approximately 0.85)

[26] One of the explanations for the high correlation is that having multiple years of prior test information helps to account for the same underlying student factors influencing achievement that are picked up by the inclusion of student background characteristics.

[27] Harris and Sass (2006) and Goldhaber et al (2012)

[28] Sass et al. (2010) find small average differences in teacher effectiveness across schools, with less affluent schools generally being staffed by less effective teachers, but also more variation in teacher effectiveness in higher poverty schools than in lower poverty schools:
Tim R. Sass, Jane Hannaway, Zeyu Xu, David Figlio, and Li Feng, "Value added of teachers in high-poverty school and lower-poverty schools," (CALDER working paper 52, November 2010).

[29] This question appears to have been the central concern in the development of the Florida value-added model, for example, which led a stakeholder committee to propose a final evaluation score that combines a teacher's individual score (measured relative to other teachers in the school) with the overall score for that teacher's school.

[30] It is also possible to use a model that makes within-student comparisons (known as a student fixed effects model). We are not aware of any states or districts considering this approach, so we have omitted discussion of this method.

[31] Teachers may not be able to address or articulate these concerns *a priori*, but it is not difficult for researchers to re-run analyses after the fact and demonstrate that rankings and decisions would have been different had a different model been used. For example, see:
Catherine S. Durso, "An analysis of the use and validity of test-based teacher evaluations reported by the Los Angeles Times," National Education Policy Center, 2012.

[32] From Goldhaber et al. (2012)

[33] For each model, teachers falling into Quintile 1 (Q1) are judged to be in the lowest 20 percent of teachers, those in Quintile 5 (Q5) in the top 20 percent.

[34] For the purposes of this discussion, we only consider models that control for prior achievement.

[35] Following Papay (2011), Goldhaber et al. (2012) report Empirical Bayes shrunken estimates and Spearman rank correlations.

[36] Also using data and results from Goldhaber et al. (2012)

[37] Note that the percentile rankings for teachers in different classrooms generated from models that include several years of prior test scores in place of student covariates look very similar to those generated from models that include only a single prior test score and student covariates (Goldhaber, 2012).

[38] These potential uses of evaluation scores raise another important issue: evaluation systems that use state exams to evaluate teachers may be less useful in these systems because scores are often not available until late in the summer, which is too late for them to be used for personnel decisions or timely professional development.

[39] Specifically, if some groups of students are more difficult to educate, and this difference is not fully captured by a student's prior performance, then models that do not account for student background will not provide accurate estimates of teacher effectiveness and will create an incentive to shy away from more difficult

classrooms. Similarly, if it is more difficult to educate students when class sizes are larger, then the same problems are present with models that do not control for class size.

[40] Another important issue is setting the "cutoff points" for teachers to be deemed high- or low-performing (known as "significance levels" in statistical parlance). When there is the possibility of negative consequences for poor evaluations, teachers would likely argue for low significance levels (and thus a very low cutoff point for a teacher to be deemed low-performing) to limit the possibility of a teacher being unfairly labeled as low-performing (known as a "Type I error"). Such a system would necessarily miss a lot of teachers who actually are low-performing (known as a "Type II error"), which would certainly hamper efforts to use the evaluations in the most effective way possible. This is yet another difficult decision that districts need to tackle.

[41] This is particularly true since the growth percentile of each student gets reported to students and families in Colorado, so SGP estimates of teacher effectiveness are consistent with what is already being distributed.

[42] Florida plans to use a single year of data.

[43] For more discussion about the relevant comparison group for teacher value-added estimates, see: Kimberlee C. Everson, Erika Feinauer, and Richard R. Sudweeks, "Rethinking Teacher Evaluation: A Conversation about Statistical Inferences and Value-Added Models." Harvard Educational Review, 83, no. 2, (Summer 2013): 349-370.

[44] It is worth noting, however, that the differences in rankings that result from the choice of model tend to be smaller than the differences in rankings that are observed for individual teachers across years or individual teachers measured according to different tests.

[45] New York State Department of Education, Guidance of New York State's annual professional performance review for teachers and principals to implement education law 3012-c and the commissioner's regulations, June 2012.

[46] Undermining trust in student growth measures would potentially limit positive teacher behavioral responses to performance feedback because they might not believe the measure is meaningful. Moreover, the issue of model shifts leading to performance estimate changes compounds what is already perceived as a problem of performance estimate reliability. For more on this see: Loeb, Susanna, and Christopher Candelaria. Carnegie Knowledge Network, "How Stable are Value-Added Estimates across Years, Subjects, and Student Groups?" Last modified October 2012. URL = <http://carnegieknowledgenetwork.org/briefs/value-added/value-added-stability/.

[47] Part of the process of making the effect of model choice transparent to stakeholders is applying different models to policy relevant teacher samples, i.e. the teachers in the states or districts that are considering adopting a student growth measure. The argument for this is that the degree to which different models will provide different performance estimates will depend not only on the relationship, for instance, between student background and test achievement but also on the extent to which students are non randomly sorted across classrooms. While the relationship between student background and achievement is unlikely to be location dependent, the sorting of students may well be influenced by other state or local policies (e.g. desegregation, the use of weighted student funding formulas, etc.).

AUTHORS

Dan Goldhaber is the Director of the Center for Education Data & Research and a Professor in Interdisciplinary Arts and Sciences at the University of Washington Bothell. He is also the co-editor of *Education Finance and Policy*, and a member of the Washington State Advisory Committee to the U.S. Commission on Civil Rights. Dan previously served as an elected member of the Alexandria City School Board from 1997-2002, and as an Associate Editor of *Economics of Education Review*. Dan's work focuses on issues of educational productivity and reform at the K-12 level, the broad array of human capital policies that influence the composition, distribution, and quality of teachers in the workforce, and connections between students' K-12 experiences and postsecondary outcomes. Topics of published work in this area include studies of the stability of value-added measures of teachers, the effects of teacher qualifications and quality on student achievement, and the impact of teacher pay structure and licensure on the teacher labor market. Previous work has covered topics such as the relative efficiency of public and private schools, and the effects of accountability systems and market competition on K-12 schooling. Dan's research has been regularly published in leading peer-reviewed economic and education journals such as: *American Economic Review, Review of Economics and Statistics, Journal of Human Resources, Journal of Policy and Management, Journal of Urban Economics, Economics of Education Review, Education Finance and Policy, Industrial and Labor Relations Review,* and *Educational Evaluation and Policy Analysis.* The findings from these articles have been covered in more widely accessible media outlets such as National Public Radio*,* the *New York Times,* the *Washington Post, USA Today,* and *Education Week*. Dr. Goldhaber holds degrees from the University of Vermont (B.A., Economics) and Cornell University (M.S. and Ph.D., Labor Economics).

Roddy Theobald is a Ph.D. student in statistics at the University of Washington, a research assistant at the Center for Education Data and Research (CEDR), and a former 7th-grade math teacher in the Oakland (CA) Unified School District. At CEDR, he has worked on projects assessing the determinants and implications of teacher layoffs, evaluating teacher training programs using student test scores, and estimating the impact of collective bargaining provisions on teacher mobility and retention. His research has been accepted for publication at *Economics of Education Review, Education Finance and Policy,* and *Education Next*, and he has served as a peer reviewer for *Education Finance and Policy* and *Educational Evaluation and Policy Analysis*.

ABOUT THE CARNEGIE KNOWLEDGE NETWORK

The Carnegie Foundation for the Advancement of Teaching has launched the Carnegie Knowledge Network, a resource that will provide impartial, authoritative, relevant, digestible, and current syntheses of the technical literature on value-added for K-12 teacher evaluation system designers. The Carnegie Knowledge Network integrates both technical knowledge and operational knowledge of teacher evaluation systems. The Foundation has brought together a distinguished group of researchers to form the *Carnegie Panel on Advancing Teaching to Improve Learning* to identify what is and is not known on the critical technical issues involved in measuring teaching effectiveness. Daniel Goldhaber, Douglas Harris, Susanna Loeb, Daniel McCaffrey, and Stephen Raudenbush have been selected to join the Carnegie Panel based on their demonstrated technical expertise in this area, their thoughtful stance toward the use of value-added methodologies, and their impartiality toward particular modeling strategies. The Carnegie Panel engaged a User Panel composed of K-12 field leaders directly involved in developing and implementing teacher evaluation systems, to assure relevance to their needs and accessibility for their use. This is the first set of knowledge briefs in a series of Carnegie Knowledge Network releases. Learn more at **carnegieknowledgenetwork.org**.

CITATION

Goldhaber, Dan, and Roddy Theobald. Carnegie Knowledge Network, "Do Different Value-Added Models Tell Us the Same Things?" Last modified November 2013. URL = <http://carnegieknowledgenetwork.org/briefs/value-added/different-growth-models/>

Carnegie Foundation
for the Advancement of Teaching

Carnegie Foundation for the Advancement of Teaching
51 Vista Lane
Stanford, California 94305
650-566-5100

Carnegie Foundation for the Advancement of Teaching seeks to vitalize more productive research and development in education. We bring scholars, practitioners, innovators, designers, and developers together to solve practical problems of schooling that diminish our nation's ability to educate all students well. We are committed to developing networks of ideas, expertise, and action aimed at improving teaching and learning and strengthening the institutions in which this occurs. Our core belief is that much more can be accomplished together than even the best of us can accomplish alone.

www.carnegiefoundation.org

We invite you to explore our website, where you will find resources relevant to our programs and publications as well as current information about our Board of Directors, funders, and staff.

Knowledge Brief 4
October 2012, Revised November 2013
carnegieknowledgenetwork.org

Funded through a cooperative agreement with the Institute for Education Science. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.