Introduction

It's hard to say anything new about value-added models (VAMs). They've been hailed as the technologically sophisticated answer to identifying the most effective teachers (Felch, Song, & Smith, 2010; Mihaly, McCaffrey, Staiger, & Lockwood, 2013; Sanders & Horn, 1994; Sanders & Rivers, 1996), and they promise to help reduce achievement gaps by enabling decision-makers to provide these teachers to our most disadvantaged children. An effective teacher, it is argued, can eliminate achievement gaps more cost-effectively than small class sizes, (Rivkin, Hanushek & Kain, 2005), perhaps even within five years (Hanushek, 2009).

Value-added models are a group of statistical models that some contend will help us find these most effective teachers. VAMs seek to measure a teacher's (or school's) performance by comparing their students' growth on standardized test scores to the average growth of other students. To do this, value-added models must isolate the contribution, or the value added by each teacher in a given year. This value added is increasingly used in evaluation to hold teachers accountable for how much their students are learning.

Along with the hyperbole, sound arguments have also been advanced regarding the usefulness of value-added approaches. For example, Doran and Lockwood (2006) offer the following rationale:

> A basic truism of learning implies that an individual student, not a student group, has increased in knowledge and skills during a specific period of time. As such, analytical methods concerned with student learning should reasonably reflect this basic principle and consider individual students as the unit of analysis with their growth trajectories employed as outcomes (p. 205).

On an instinctive level, the approach makes sense. We care about student learning, and we ought to attempt to measure it on the basis of how much each individual student is growing from one time-point to another. However, VAMs have also been criticized on several fronts. To wit, they are unstable from year to year: only 35% of teachers ranked in the top fifth on teacher value-added measures one year were still ranked in the top fifth in the next year (Koedel & Betts, 2007). They are imprecise: even with three years of data, the margin of error for a teacher at the 43rd percentile ranges from the 15th to the 71st (Corcoran, 2010). They are inaccurate: Type I (false positive) and Type II (false negative) error rates for comparing a teacher's performance with the average are likely to be about 25% with three years of data (Schochet & Chiang 2010). Estimates differ greatly depending on what test is used as the outcome (Lockwood, McCaffrey, Hamilton, Stecher, Le & Martinez, 2007a). Causal interpretations of the estimates are questionable, and they implicitly assume

randomly equivalent groups—an assumption that is not justified given the ample evidence demonstrating the non-random sorting of students to schools and teachers. In other words, students are intentionally sorted into specific teachers' classrooms, perhaps because one teacher is considered good with students with behavior problems, or because another teacher speaks Spanish, and students, and students are systematically grouped into different schools based on socioeconomic factors, segregated housing patterns, and parental preference, among other factors. Rothstein (2010) demonstrated some of the problems with causal interpretations and provided evidence of non-random sorting by showing backward effects, such as fifth grade teachers shown to have large "effects" on fourth-grade achievement—a technical impossibility.

In spite of these criticisms, value-added modeling (VAM) is increasingly becoming an entrenched component of teacher evaluation in the U.S. In 2009, the New Teacher Project released The Widget Effect, a report sharply criticizing an education system in which "all teachers are rated good or great" (Weisberg, Sexton, Mulhern, & Keeling, 2009, p. 6). But things have changed substantially since then. In just the past few years, many district and state teacher assessment frameworks, including New York City, Los Angeles, Chicago, Washington, D.C., Florida, Ohio, and Arizona among others, have adopted teacher evaluation systems that base one-third to one-half of a teacher's rating on the achievement gains of their students—and more appear to be moving in this direction. Research, including most notably the Measures of Effective Teaching (MET) Project (an expansive multi-year study of thousands of teachers funded by the Bill and Melinda Gates Foundation), has also featured some prominent advocacy for VAMs (MET Project, 2013). Yet, among the most passionate critics and defenders of these statistical instruments, only a very few understand the underlying models. Technical explanations are typically opaque; accessible reports usually skim the surface or repeat one another. There is not even general agreement about the acronym VAM itself: value-added models is typical, but value-added measurement is used almost as frequently, and many writers choose the more general VA.[1]

The "s" in VAMs is important because value-added modeling encompasses many different models, and there is little public awareness or discussion of the differences among them. These models differ on such basic questions as, "Should past scores be used to predict current scores, or to determine the difference between past and current scores?", "Should the assumption be that students learn at a constant rate?", and "Do the effects of teachers persist indefinitely into the future or diminish over time?" The

[1] The models are also referred to as value-added assessment or value-added analysis, so the acronym VAA is also commonly, and inconsistently, used.

conceptual differences among models are important, and arguments can be made in favor of one model or another as more conceptually valid, but, rather than engage in those arguments, we simply aim to clarify the mathematics and the assumptions underlying different approaches and provide evidence as to the extent to which those differences matter.

In this paper, we seek to provide a coherent overview of the main value-added models and sketch out the implications of each model for the inferences drawn about individual teachers. In doing so, we hope to help bridge the gap between practitioners and researchers. Practitioners will gain understanding of these complex models; researchers will gain insight into the potential consequences of these technical decisions.

Our explanations of the different models draw heavily from overviews by Wiley (2006); McCaffrey, Lockwood, Koretz & Hamilton (2003); Green (2010); Sanders & Horn (1994); Raudenbush & Bryk (2002); and Lockwood, et al. (2007a). Existing VAM research often compares models in terms of the overall correlations of value-added estimates they produce. However, because these models are increasingly used to make norm-referenced inferences about individual teachers and applied for high-stakes purposes (i.e., decisions such as promotion, compensation, or dismissal) we examine the implications of utilizing different VAMs on the relative ranking of teachers. A relatively high correlation between two models could still mean that a substantial number of individual teachers would receive very different estimates of effectiveness under the two models. Therefore, we focus on clarifying the impacts of model choice for individual teacher rankings rather than on overall correlations.

Methods

Data

The dataset we used in this analysis was the annual state mathematics achievement scores of a cohort of elementary school students (N=9,295) from a large urban school district. Students were linked to teachers as they progressed from first through fifth grade in years 1997–1998 through 2001–2002. The variables are student ID, teacher ID, test score, and grade level (scaled from 0 – 4). We have no data on students' race, socioeconomic status, or other covariates. For our purposes, this simple dataset helps to illustrate the many different modeling approaches that are possible, though it also means that we leave aside the important question of which covariates should be included (see Lockwood, et al., 2007a, for a discussion of this question). The data were previously explored by Lockwood, McCaffrey, Mariano, and Setodji (2007b) and Mariano, McCaffrey, and Lockwood (2010). The original researchers rescaled the test

scores to produce marginal means ranging from roughly 3.5 to 6.2 and marginal standard deviations ranging from roughly 0.92 to 1.08 across grades. Also, linkages between students and teachers were only available for 54% of the records due to student movement into and out of the district. As is commonly the case with student achievement data, the vast majority (79%) of students had at least one missing year of test score data over the 5 years. Thus, for simplicity, we restricted our analysis to only those students with both fourth and fifth grade testing data and whose records could be clearly linked to a teacher for each year. Our working sample was 3,661 students and 257 fifth grade teachers.

Analytic Approach

Though VAM is becoming commonplace in teacher evaluation systems, there is no consensus on the most accurate model. In fact, it is probably not even appropriate to speak of more and less accurate models, since the trait that is being measured, teacher effectiveness, is not completely defined or agreed-upon. Trade-offs must be made in terms of the simplicity and comprehensibility of a model versus its precision, as well as in the assumptions underlying different models. Wiley (2006) explains the defining features of a value-added model: "1) it studies change in the performance of individual students, and 2) it seeks to determine to what extent changes in student performance may be attributed to particular schools and teachers" (p. 6). This broad definition encompasses many different models, which we group into three classes: *gain score models*, *covariate adjusted models*, and *multivariate models*. As we explain each model, we use the dataset described above to exemplify the properties of these different types of models in terms of the assumptions each model requires and the extent to which the results differ from model to model.

Gain Score Model

Table 1. Gain Score for Teacher #92 and Teacher #116

Teacher	Student	5th Grade Math Score	4th Grade Math Score	4th to 5th Grade Gain	Teacher Average Gain	Total Average Gain	Teacher Effect (Gain Score)
	1	6.93	5.60	1.33			
	2	6.40	5.43	0.98			
92	4	7.50	5.95	1.55	1.30	0.54	0.76
	5	5.68	4.20	1.48			
	6	7.03	5.78	1.25			
	7	6.00	5.78	0.23			
116	8	5.08	5.15	-0.08	0.30	0.54	-0.24
	9	6.58	5.83	0.75			

Note: Table adapted from A Practitioner's Guide to Value Added Assessment (Wiley, 2006).

The gain score model is the simplest VAM with a basic gain score formed for each student by subtracting the student's prior test score from her current score. The gain score model is single-wave (test scores collected over two years only) and univariate. Students with missing data are excluded from the gain score model. A teacher's gain score (i.e. their teacher effect) can be found using the following equation

$$y_t - y_{t-1} = m_t + T_t \quad (1)$$

where *yt* denotes student score at time *t*; *yt-1* denotes student score in prior year; *mt* denotes student specific mean gain; and *Tt* denotes teacher effect.

Table 1 shows how the gain score was calculated for two sample teachers in the dataset. A teacher's gain score equals the average gain of all of a teacher's students minus the total average gain of all teachers in the dataset (in our case 0.54). Thus, Teacher #92 had a gain score of 1.30 -0.54 = 0.76 and Teacher #116 a gain score of 0.30 - 0.54 = -0.24. Notice that while Teacher #116 had a positive average gain overall of 0.30, because her gain was below average, she had a negative gain score. More complex variants of the gain score model add school (*St*) and district (*Dt*) effects to further tease out achievement.

Covariate Adjusted Model

Often VAMs try to account for student background characteristics by including covariates in the model such as student socioeconomic status or minority status. In our data, the only covariate being used to adjust the estimates is the students' prior year test scores. Like the gain score model, the covariate adjusted model takes into account only the current and prior year's scores, so these models are often referred to as single-wave or univariate. Again, students with missing data are excluded. Unlike the gain score model, which subtracts the previous year's score, in the covariate adjusted model the effect of the previous year's score is used as a covariate to adjust the estimate, like an analysis of covariance (ANCOVA). The difference between the covariate adjusted and a typical ANCOVA is that the slope is not assumed to be equal across classrooms; instead the model estimates the slope for each classroom (teacher). The covariate adjusted model is essentially a regression model with the prior year's score as the primary covariate and a dummy variable for each teacher. A teacher's current score is specified as a function of student's prior score and possibly other covariates. Teacher average gain is then compared to the total gain to produce the covariate adjusted teacher effect. The covariate adjusted model has been used in Dallas Schools (DVAAS) with student and school level covariates. A teacher's covariate adjusted score can be found using the following equation:

$$y_t - by_{t-1} = m_t + T_t + e_t \quad (2)$$

where *yt* denotes student score at time *t*; *byt-1* denotes student score in prior year adjusted by any covariates; *mt* denotes student specific mean gain; *Tt* denotes covariate adjusted teacher effect; and *et* denotes residual errors

Table 2. Covariate Adjusted Calculation for Teacher #92 and Teacher #116

Teacher	Student	5th Grade Math Score	4th Grade Math Score	4th to 5th Grade Gain	Teacher Average Gain	Total Average Gain	Teacher Effect (Covariate Adjusted Score)
92	1	6.93	5.78	1.15	1.22	0.68	0.54
	2	6.40	5.20	1.20			
	3	6.53	5.29	1.24			
	4	7.50	6.47	1.03			
	5	5.68	4.08	1.60			
	6	7.03	5.94	1.09			
116	7	6.00	5.57	0.43	0.49	0.68	-0.19
	8	5.08	5.44	-0.36			
	9	6.58	6.16	0.42			

Note: Table adapted from A Practitioner's Guide to Value Added Assessment (Wiley, 2006).

Table 2 displays how the covariate adjusted teacher effect was calculated for the same two teachers. Notice that in this model students' prior year's score has been adjusted. Teacher #92 now has a slightly lower teacher effect as estimated by the covariate adjusted model (1.22 - 0.68 = 0.54) compared to the gain score model (0.76); Teacher #116 now has a slightly higher covariate adjusted teacher effect, -0.19 vs. -0.24, when compared with the gain score model. Though these teacher effect estimates differ only slightly between models, as we will see, for some teachers the difference can be quite stark.

Note that it is also possible to include a random term at level 2 (uoj) so as to allow the intercepts to vary. Introducing a random effect transforms the gain score model to an ANOVA model with random effects and the covariate adjusted model to an ANCOVA model with random effects (see Table 3). The inclusion of a random effect into the model results in individual teacher effect estimates being empirically Bayes (EB) adjusted. The adjustment is based on classroom size and the distance each classroom's mean achievement is from the overall mean achievement. A more complete discussion of ordinary least squares (OLS) vs. EB effect estimates occurs in the second section under Value-Added Model Assumptions, Fixed vs. Random Teacher Effects.

Table 3. Adding Random Effects (uoj) to Gain Score and Covariate Adjusted Models

Gain Score Model → Random Effects ANOVA Model	Covariate Adjusted Model → Random Effects ANCOVA Model
Level 1: Student: $Y_{it} = \beta_{oj} + e_{ij}$	Level 1: Student: $Y_{it} = \beta_{oj} + \beta_{1j} X_{ij} + e_{ij}$
Level 2: Teacher: $\beta_{oj} = \gamma_{oo} + u_{oj}$	Level 2: Teacher: $\beta_{oj} = \gamma_{oo} + u_{oj}$ $\beta_{1j} = \gamma_{10}$
Combined: $Y_{it} = \gamma_{oo} + u_{oj} + e_{ij}$ where:	Combined: $Y_{it} = \gamma_{oo} + \gamma_{1o} X_{ij} + u_{oj} + e_{ij}$ where:
Y_{it} denotes student gain score	Y_{it} denotes student 5th grade score
β_{oj} is a random intercept associated with jth teacher	β_{oj} is a random intercept associated with jth teacher
e_{ij} is random error	β_{1j} is a random slope associated with the jth teacher
γ_{oo} is the mean of random intercept	X_{it} denotes student 4th grade score
u_{oj} is the random effect of the jth teacher on the random intercept	e_{ij} is the random error
	γ_{oo} is the mean of random intercept
	γ_{1o} is the mean of random slope
	u_{oj} is the random effect of the jth teacher on the random intercept

Multivariate Layered Model

The layered model is the oldest multivariate value-added model. Created by Dr. William Sanders in 1992, the layered model is still used today in the Tennessee Value-Added Assessment System (TVAAS). Unlike the gain score and covariate adjusted models, the layered model utilizes test scores from multiple years and can estimate scores for missing students. As the name implies, later years of teacher effects build upon the gains from earlier years. Student growth is not assumed constant over time and teacher effects persist undiminished. The layered model does not include student covariates or school effects. To use the developers' own words, "TVAAS uses a highly parsimonious model that omits controls for SES, demographic, or other factors that influence achievement" (Ballou, Sanders & Wright, 2003, p. 60, as cited in Wiley, 2006, p. 28). And, "Each student serves as his or her own control, creating a level playing field and eliminating the need to adjust for race, poverty, or other socioeconomic factors" (Schooling Effectiveness, SAS® EVAAS® for K-12, http://www.sas.com/en_us/industry/k-12-education/evaas.html, as cited in Wiley, 2006, p. 29).

A simplified version of the layered model as presented by McCaffrey et al. (2003) for three years of testing is shown below.

$$yi1 = m1 + T1 + ei1 \quad (3)$$
$$yi2 = m2 + T2 + T1 + ei2$$
$$yi3 = m3 + T3 + T2 + T1 + ei3$$

In the layered model, student *i*'s achievement score at time *t* is a function of the mean achievement score of the overall student population at time *t* (*mt*), the student's teacher *T* at time *t* and all previous times (*Tt, Tt-1, Tt-2, Tt-3*….), and a residual error term at time *t* (*eit*).

Multivariate Cross-Classified Model

The cross-classified model is also a multivariate, multiple-wave model with a two-way cross-classification of repeated measures and teachers at level 2 (Raudenbush & Bryk, 2002). In this model grades are nested within students, grades are nested within teachers, and students and teachers are crossed because students have different teachers in each grade as shown in Figure 1 below. Each year, as the crossed arrows indicate, students change classrooms and thus the nested structure changes each year.

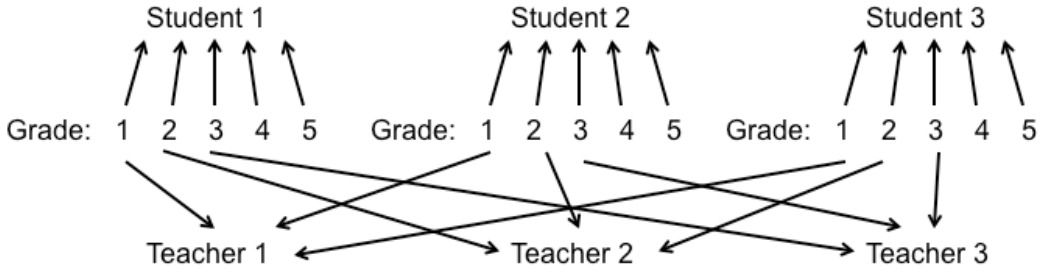

Figure 1. The Cross-Classified Value-Added Model

Unique to this model, individual student growth is assumed to increase at a constant linear rate over time. Therefore, due to this assumption, teachers with students with high growth rates in previous years will tend to have lower cross-classified VAM scores than those estimated in other models that make no assumption of constant academic growth (e.g. layered model). A simplified version of the cross-classified model, as presented by McCaffrey et al. (2003), is shown below.

$$yi1 = mi + bi + T1 + ei1 \quad (4)$$
$$yi2 = mi + 2bi + T1 + T2 + ei2$$
$$yi3 = mi + 3bi + T1 + T2 + T3 + ei3$$

In the cross-classified model, student *i*'s achievement score at time *t* is a function of the student's mean initial achievement score (*mi*), the student's linear achievement growth slope (*bi*), the student's teacher at time *t* (*Tt*), and a residual error term at time *t* (*eit*). Therefore, student *i*'s growth is given as *mi* + *bit* with the teacher's effect the permanent deflections from this growth curve (Raudenbush & Bryk, 2002).

Multivariate Variable Persistence Model

Both the layered and cross-classified multiyear models make a very strong assumption—student learning gains produced by any one teacher are presumed to persist undiminished into the future. That is, the models assume a particular 2nd grade teacher has the same impact on her students' 2nd grade assessment as she does on the same students' 5th grade assessment. However, any decay in teacher effects directly impacts future value-added estimates. The variable persistence model attempts to account for this deterioration in teacher effects on student achievement over time by weighting/scaling the effects by a persistence parameter (McCaffrey et al., 2004). A simplified version of the variable persistence model, presented in McCaffrey et al. (2003), is shown below.

$$yi1 = m1 + T1 + ei1 \quad (5)$$
$$yi2 = m2 + T2 + a21T1 + ei2$$
$$yi3 = m3 + T3 + a32T2 + a31T1 + ei3$$

Notice the variable persistence model is identical to the layered model, but with the addition of this persistence parameter (*a*) in later years. Thus, student *i*'s achievement score at time *t* is a function of the mean achievement score of the overall student population at time *t* (*mt*); the student's teacher at time *t*; the student's teacher at all previous times (*Tt-1*, *Tt-2*, *Tt-3*….)—but these teacher effects are all adjusted by this persistence parameter (*a*) and a residual error term at time *t* (*eit*).

Value-Added Model Assumptions

Lord's Paradox

These models differ most in the assumptions they make about how teachers influence student learning. For example, Lord (1969) showed that two

perfectly reasonable researchers can come to very different conclusions analyzing the same dataset, depending on the research question asked. This irresolvable contradiction, now known as Lord's Paradox, can occur when comparing value-added estimates across models, as neither the gain score model nor the covariate adjusted model is more correct, but ask different questions.[2] While the gain score model (ANOVA) "asks" how initial scores differ from final scores, the covariate adjusted model (ANCOVA) "asks" how one variable can predict variation in another. In other words, how can achievement at the beginning (4th grade scores) be used to predict achievement (5th grade scores) at the end? At times these different questions can produce different answers.

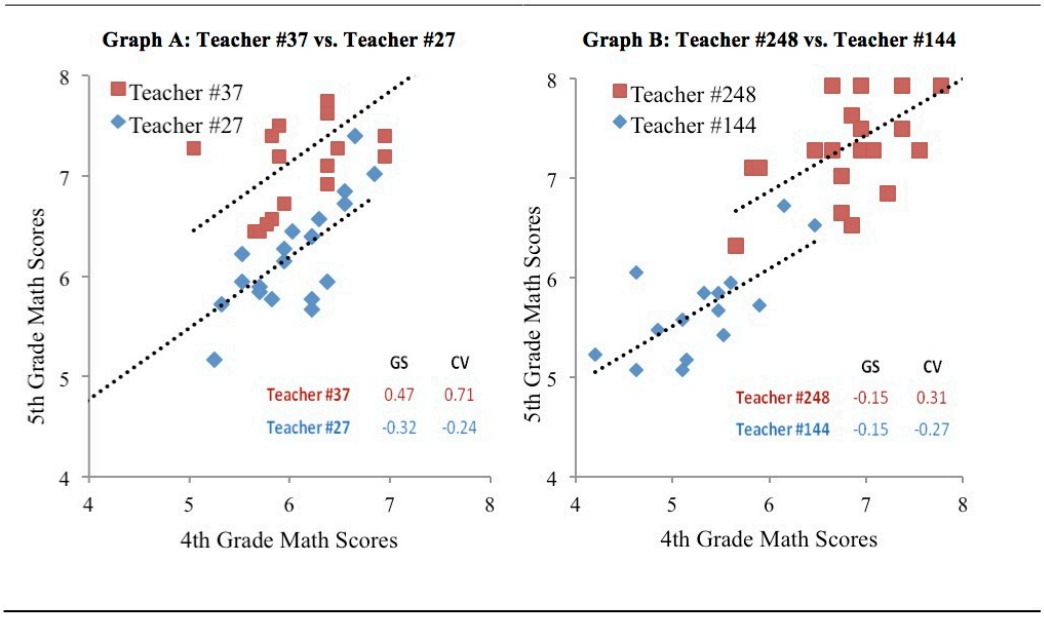

Figure 2. Math 4th and 5th Grade Student Achievement Scores, by Teacher's Classroom

Graphs A and B in Figure 2 show Lord's Paradox in action using our dataset. The scatterplots compare 4th and 5th grade math achievement scores for students taught by four different teachers. Graph A displays the more common situation when the gain score model and covariate adjusted model produce similar teacher effect estimates. In Graph A, Teacher #37 has a higher gain score than Teacher #27 (0.47 vs. -0.32) as well as a higher covariate adjusted score (0.71 vs.

[2] For a very accessible, two-page explanation of Lord's Paradox see *A Paradox in the Interpretation of Group Comparisons* by Frederick Lord (see Bibliography).

-0.24). Both models agree Teacher #37 should be credited with greater student test score gains.

Graph B, however, shows the models can at times produce contradictory teacher effects. Teacher #248 has an unambiguously higher covariate adjusted score than Teacher #144 (0.31 vs. -0.27), yet both teachers have the same estimated gain score of -0.15. Why then the curious results in Graph B? Teacher #248 is teaching a classroom of especially high achieving students and, therefore, there is little overlap in the distributions of student test scores between these teachers. Looking at only these two extreme cases is, of course, overstating the point because these two teachers do overlap with many other teachers in the sample, even though they don't overlap with one another. However, it helps to illustrate the well-known problem that no statistical procedure can fully account for uncontrolled pre-existing classroom differences.

Fixed vs. Random Teacher Effects

There is also the question of fixed vs. random teacher effects. Treating teacher effects as fixed assumes observed teachers are the only teachers of interest; treating teacher effects as random assumes teachers are drawn from a larger population. How one chooses to handle this assumption changes the rank order of individual teachers. Assuming fixed effects results in estimated teacher effects that depend only on the teacher's students, possibly adjusted for background covariates (e.g., gender, race, and SES). In contrast, random teacher effects are calculated using estimated best linear unbiased predictors (EBLUP's) or empirical Bayes (EB) estimators (Raudenbush & Bryk, 2002). Unlike ordinary least squares (OLS) estimators, EB estimators use data (i.e. borrow strength) from all teachers in the sample to estimate each teacher's value-added effect.

Table 4. Ten Teachers' EB and OLS Teacher Effect Scores and Shrinkage Estimates

Teacher #	Sample Size (n)	Teacher Average Gain (OLS)	Shrinkage	Teacher Average Gain (EB)	Grand Mean (γ)	EB Gain Score
36	4	0.04	0.17	0.21	0.54	-0.33
60	7	0.01	0.12	0.13	0.54	-0.41
116	3	0.30	0.10	0.40	0.54	-0.14
237	12	0.32	0.03	0.35	0.54	-0.19
4	12	0.38	0.02	0.40	0.54	-0.14
249	21	0.84	-0.03	0.82	0.54	0.28
14	14	0.97	-0.05	0.91	0.54	0.37
130	18	1.33	-0.08	1.24	0.54	0.71
92	6	1.30	-0.19	1.10	0.54	0.56
197	13	2.83	-0.31	2.51	0.54	1.97

Table 4 displays the EB adjusted gain scores of 10 teachers.[3] A teacher's EB average gain is calculated by adding the teacher's original OLS average gain to their estimated shrinkage. Teachers with OLS average gains farther away from the grand mean and in classrooms with fewer students will show greater shrinkage. Notice Teachers #116 and #92 introduced earlier in Table 1 and Table 2. Teacher #116 had an average gain of 0.30 and only has three students. Mostly as a result of her small classroom size, her OLS estimated average gain of 0.30 was shrunk upward to 0.40, and closer to the grand mean of 0.54. In contrast, Teacher #92's OLS estimated average gain estimate of 1.30 was shrunk downward to 1.10. Though Teacher #92 had more students (six vs. three), her OLS estimated average gain of 1.30 exhibited greater shrinkage because it was far away from the overall grand mean (0.54), and thus more extreme. In order to center the gain at zero, grand mean gain is subtracted from a teacher's EB average gain to determine a teacher's EB gain score.

Shrinking OLS gain scores by including a random effect reduces the variance of the teacher effect estimate relative to the fixed-effect estimate. However, the obvious drawback of including a random effect is, by definition, shrinking OLS scores introduces bias. For highly effective teachers, EB gain score estimates can be far below the teacher's gain score; for highly ineffective teachers, EB gain score estimates can be far above the teacher's gain score. This is particularly true if the teacher's class is small. Consequently, if accountability decisions are made on the basis of teachers with extreme scores, teachers with smaller classes will tend to be excluded using EB adjusted gain scores (rightly or wrongly). Of course, because the fixed-effect gain score for teachers with small classes will more likely fall in the extremes of the distribution, using fixed effect scores will result in high-stakes decisions being applied more often to teachers with small classes and in a more erratic fashion as extreme values will be driven in part by random fluctuations. Given these trade-offs, McCaffrey et al. (2003) offer the following guidelines:

> Random-effects models are preferred when estimates that shrink teachers toward the mean—possibly underestimating the most and least effective teachers—are less detrimental to the inference of interest than estimates with large but unsystematic errors. Fixed-effects models are preferred otherwise (p. 67).

[3] Though Table 4 only reports EB gain score estimates, a random effect was also included in the covariate adjusted model and a similar degree of shrinkage was found.

Tekwe et al. (2004) recommend using fixed effects models for simplicity and ease of interpretation; however, more recent VAMs have utilized random effects almost exclusively (McCaffrey et al., 2003).

School and District Effects

School and district inputs, such as policies regarding curricula, interventions, and other academic programs, clearly exert considerable influence on student achievement. Whether one chooses to account for these factors in VAMs can also drastically influence teacher effect estimates. Although students are non-randomly sorted across schools and districts, these fixed effects are often omitted in current value-added models including TVAAS. Like the decision concerning the inclusion of random effects, there is an inherent trade-off. Omitting school (district) effects can bias "true" teacher effects through mistaken attribution (i.e., a teacher could receive heavier blame/credit than is warranted). However, adding school or district fixed effects can limit the usefulness of the results as their inclusion restricts comparison to only those other teachers in the same school or district (Baker et al., 2010). Given that teachers do not randomly sort across schools and districts, this can result in an average teacher being unfairly punished if surrounded by above average teaching colleagues, or rewarded if surrounded by below average colleagues. For instance, a teacher in a school with exceptionally talented teachers may not appear to add as much value to her students as others in the school, but if compared to all district teachers, might fall well above average. That said, effective teachers may also make their fellow teachers at the school more effective by establishing good study patterns and discipline in their students that perhaps rubs off in other classes. It is impossible to disentangle these effects. It comes as no surprise then that McCaffrey et al. (2003) found the decision as to whether to include school fixed effects significantly changed value-added inferences.

Rate of Persistence

An often cited claim of proponents of VAM is that matching disadvantaged students with high value-added teachers for five years in a row would be enough to completely eliminate the achievement gaps between black and white or poor and non-poor students (Hanushek, 2009). However, this assertion is based on the assumption of perfect persistence of teacher effects, whereas Jacob, Lefgren, and Sims (2010) have found that as little as one-fifth to one-third of teacher-induced learning gains persist into the following year. To illustrate how varying the degree of persistence can strongly influence teacher effect estimates, consider a class of 5th grade students with an average

achievement score of 100 at the end of third grade and assume true teacher effects diminish by 20% as shown in Table 5.

Table 5. Undiminishing vs. Diminishing Teacher Effect Estimates on Future Value-Added Estimates

Undiminished		Diminished	
100	Classroom average achievement at the end of 3rd grade	100	Classroom average achievement at the end of 3rd grade
+10	Contribution of 4th-grade teacher to 4th-grade test; the original teacher effect of 10 has continued undiminished over the single year	+8	Contribution of 4th-grade teacher to 4th grade test; the original teacher effect of 10 is assumed to have diminished by 20% over the single year
115	Average score on 5th grade test, then	115	Average score on 5th grade test, then
+5	Attributed 5th grade teacher effect	+7	Attributed 5th grade teacher effect

Note: Table adapted from A Practitioner's Guide to Value Added Assessment (Wiley, 2006).

If the "true" rate of decay in teacher effects is 20%, the 4th grade teacher in the undiminished model is wrongfully given credit for a portion of student achievement gains (+2 points) that should have been attributed to the 5th grade teacher. Thus, the effects of the 5th grade teacher are underestimated by two points in the undiminished model. Similarly, had we incorrectly assumed teacher effects diminished by 40%, the effect of the 5th grade teacher would have been overestimated by two points in the diminished model. Rate of persistence over time in teacher effects is highly controversial for good reason. Changing the strength of the rate of persistence can drastically alter teacher effect estimates; yet it is an exceedingly complex parameter to measure.

Jacob et al. (2010) attempted to tackle this conundrum of estimating teacher effect estimates. After breaking knowledge into its long-run and short-run components, the authors found test score variation as estimated by teacher value-added models is only about 20% as persistent as true long-run knowledge. The authors argue this low persistence in variation of teacher quality is not due to value-added measures, but is common to several other methods of measuring teacher quality. When assertions are made that a high value-added teacher can raise math achievement by one-fifth of an "average yearly gain" (Aaronson, Barrow, & Sander 2007), these claims are made under the assumption of perfect persistence; however, the long run gain may be closer to 0.04, and thus a teacher's true ability may be overstated.

Nonetheless, while the assumption of complete persistence of teacher effects seems unlikely to hold and has been shown to be false in some empirical

data (Jacob et al., 2010), the variable persistence model that relaxes that assumption is itself problematic. The variable persistence model relies on a mixture of gains and statuses of students (McCaffrey, 2010) and thus, is likely to entangle teacher effect estimates with the academic and SES level of those students taught. Unlike the evidence that teacher effects decay, in many cases, we have no sound empirical evidence preferring one set of assumptions to another. Moreover, in some important instances, the empirical evidence points in opposing directions, such as McCaffrey's (2010) findings that a model that assumed complete persistence of teacher effects had larger errors than the variable persistence model, but that the errors in the variable persistence model were correlated with student background characteristics and thus, might be systematically biased. Therefore, it may be impossible to definitively determine which model is "best" and any estimates of teacher effects produced by one model could arguably be said to have a "margin of error" encompassing not simply the typical confidence interval but the range of the estimates produced for that teacher by all other models.

If we cannot determine which model is "best," then it makes sense to look carefully at how big the differences are between models in terms of practical importance. We turn to this question in the next section.

Comparing the Results from Different Models

The teacher effect estimates from the gain score, covariate adjusted, cross-classified, and variable persistence VAMs all correlated very highly. Only the variable persistence model—the only model to account for diminishing teacher effects—did not correlate above 0.90 with the other models (see Table 6).

Table 6. Correlation of Teacher Effect Estimates Between VAMs

	Gain Score	Covariate Adjusted	Cross-Classified	Variable Persistence
Gain Score	1.00			
Covariate Adjusted	0.91	1.00		
Cross-Classified	0.96	0.93	1.00	
Variable Persistence	0.81	0.96	0.86	1.00

The competing models also produce similar aggregate student 4th grade to 5th grade growth estimates (see Table 7). The EB and OLS gain score models calculated one-year gain at 0.54, while the OLS covariate adjusted model

estimated slightly greater student gain at 0.77. The cross-classified model estimated similar one-year gain at 0.65. Taken together, the high correlations and similar fixed effects estimates across models provide evidence that in the aggregate the models produce similar growth estimates.[4]

Table 7. Estimated Aggregate Student Achievement Gain: 4th Grade to 5th Grade

	Gain	Standard Error
Gain Score	0.54	0.011
Covariate Adjusted	0.77	0.011
Cross Classified	0.65	0.010

After considering overall growth differences between the value-added models, we next examined differences in individual teacher effects. First, we ranked the 257 5th grade teachers according to their respective gain model scores and covariate adjusted model scores. While the median difference was nine percentile points between the two models, 12% of teachers had a rank difference greater than 20 percentile points. In addition, 25% of the teachers who ranked in the bottom quintile—those that might be categorized as "struggling"—in the gain score model were not ranked in the bottom quintile of the covariate adjusted model. Teacher #180 saw the greatest change in rank between models—ranked in the 6th percentile by the gain score model, yet the 70th percentile by the covariate adjusted model. Likewise, Teacher #248 (introduced earlier in the Lord's paradox example with high achieving students) was ranked in the 37th percentile by the gain score model, but in the 83rd percentile by the covariate adjusted model.

Ranking teachers by the multiyear cross-classified and variable persistence models generated comparable, if not more divergent findings. The majority of teachers (53%) had a difference of at least 10 percentile points between the two models while 23% of teachers had a rank difference greater than 20 percentile points. Moreover, 35% of teachers categorized as "struggling" by the cross-classified model—by being ranked in the bottom quintile—were not ranked in the bottom quintile of the variable persistence model. Again, Teacher #180's rank changed wildly between models—ranked in the 11th percentile by cross-classified model yet 83rd percentile in the variable persistence model. Similarly, Teacher

[4] We were unable to calculate the overall student achievement gain and its associated standard error for the variable persistence model. This reflects our limited understanding of this complex model and is not a shortcoming of the model itself.

#248's relative standing was the 46th percentile in the cross-classified model, but the 89th percentile in the variable persistence model.

However, perhaps most troubling was the shuffling of teachers in the tails of the distribution. Only half of the bottom 5% of teachers was the same across the cross-classified and variable persistence models; only 43% of the top 5% of teachers was the same across models. These results suggest that while the models behaved similarly in the aggregate, for some individual teachers model selection made an enormous difference on their perceived performance.

Standard Errors

It is possible, however, that we are overstating the case for the lack of precision in individual teacher effects. Perhaps this lack of precision is already built into these model estimates in terms of their estimated standard errors. In that case, it might be advisable to simply pay more attention to the standard errors and associated confidence intervals around each effect estimate. Estimating a series of different models may be redundant as far as providing us with a sense of the extent of uncertainty around each effect estimate. To explore this possibility, Table 8 displays the smallest, largest, and mean standard errors for two models.

Table 8. Standard Errors for Covariate Adjusted and Variable Persistence Models

Covariate Adjusted Model					Variable Persistence Model			
Teacher ID	Effect Estimate	Standard Error	95% Confidence Interval		Effect Estimate	Standard Error	95% Confidence Interval	
256	-0.48	0.08	-0.63	-0.32	-0.41	0.07	-0.55	-0.27
257	-0.52	0.46	-1.45	0.41	-0.22	0.31	-0.83	0.39
Mean	---	0.14	-0.28	0.28	---	0.13	-0.26	0.26

One way to put these standard errors in perspective is to compare them to the overall distribution of scores. The standard deviation of the distribution of teachers for the covariate adjusted model was 0.41, meaning that, if we assume a normal distribution, approximately 95% of the estimates fell between -0.82 and 0.82. The confidence interval for a teacher at the mean spans from -0.28 to 0.28, just over one-third of that distribution, while the largest confidence interval, for Teacher #257, spanned a range even larger than this overall distribution, suggesting that for Teacher #257, the model provides nothing but a guess as to her effectiveness. Teacher #256, on the other hand, with the smallest standard error, has a confidence interval that is just less than one-fifth the range of the overall 95% distribution.

For the variable persistence model estimates, the standard deviation of the distribution of teachers was 0.40, so 95% of the population ranged between about -0.80 to 0.80. The confidence interval for the mean teacher in that model ran from -0.26 to 0.26, again approximately one-third of the overall 95% distribution. The largest standard error, for Teacher #257, resulted in a confidence interval that was just over three-fourths as large as the overall distribution, and the smallest confidence interval was again less than one-fifth of the overall distribution. The standard errors are somewhat smaller for the variable persistence model, most likely because this model included multiple years of student achievement data. However, overall, these standard error estimates reinforce the lack of precision in model estimation earlier emphasized through our comparisons of the implications of various models for individual teacher effect estimates. Our findings show that model selection can have notable and even severe consequences for individual teacher effect estimates, in spite of aggregate model correlations that tend to be at 0.90 and above.

Discussion

One of the leading proponents of VAM, Eric Hanushek (2011), has suggested that replacing the lowest performing 5% to 8% of teachers based on valued-added scores with "average teachers" would increase total U.S. economic output to the tune of $112 trillion in terms of the present value of future additions to Gross Domestic Product. However, who exactly are these bottom 5% to 8% of teachers? Recall that in our findings, only half of the bottom 5% of teachers remained the same across the cross-classified and variable persistence models. Our models do not even come close to a clear-cut consensus on the bottom 5%, and they are barebones without the inclusion of divergent student background covariates.

In addition to the lack of consensus that we found based on different models, there are a host of other methodological choices that we did not consider. In fact, in surveying the literature, model selection appears to be among the least consequential choices that must be made when estimating a value-added model. For instance, Lockwood et al. (2007a) found that the choices of whether to include covariates such as student demographic variables and class mean characteristics, had nearly as large of an impact on the aggregate correlations as did the choice of model; they also found the choice of outcome variable (i.e., what test was used as the indicator of student learning) had a far greater impact than either model choice or decisions about what covariates to include. In fact, when estimating the same model utilizing different student achievement tests as the outcome variable, the authors found that the correlations among teacher effect estimates were "uniformly low", ranging from 0.01 to 0.46 depending on year,

model, and controls for student background characteristics. Sass (2008) also found large differences in estimates depending on what test was used; in a dataset of Florida teachers, he found only 43% of teachers ranked in the top 20% on the Stanford Achievement Test were also ranked in the top 20% when using the state test. Accordingly, these studies suggest that the selection of student achievement tests contributes more to the variation in how teachers are ranked than does the selection of the statistical model. In other words, if the differences we found from one model to the next are worrisome, then the findings from Lockwood et al. and Sass multiply those concerns.

Looking at a different source of inconsistency, researchers have found that even when using the same model and the same outcome variable (i.e., student achievement test), the estimates are extremely unstable from one year to the next: recall that Koedel and Betts (2007) found that only 35 percent of teachers ranked in the top fifth on teacher value-added measures one year were still ranked in the top-fifth in the next year. McCaffrey, Sass, Lockwood, and Mihaly (2009) found similar instability from one year to the next, and both papers estimated that about 10 to 15 percent of teachers fall all the way from the top fifth of teachers to the bottom fifth, while a similar number move all the way from the bottom fifth to the top fifth in just one year. In addition, it should be noted that in this paper we have not touched upon a number of other issues that researchers have raised with respect to the reliability and validity of VAM estimates: demonstrable non-random assignment of students (Rothstein, 2010); data that is missing not at random; unintended consequences such as narrowing of the curriculum; and perverse incentives that might encourage cheating, discourage collaboration, and discourage attention to migrant students who won't be included in VAM estimates.

It doesn't take much experience with the American legal system to forecast the day that moderately imaginative lawyers begin using the limitations of VAMs to sue school districts and prevent the firing of "the bottom 5%." Districts are rushing to implement reforms that use VAM as one-third to one-half of the final evaluation, and typically rely on principal observations for the other major portion of the final score. Yet imagine this argument in court for a wrongful discharge case: "But your honor, we have shown how the principal of this school began to record low evaluations for Mr. Smith at the same time Mr. Smith began to become politically active in the union, and furthermore, Mr. Smith's supposedly low value-added estimate jumps to the 28th percentile when estimated using a different model…" Many districts plan to include student surveys in a multiple measure evaluation, but survey research suggests that students tend to rate their teachers uniformly high (Follman, 1992), so survey results might not provide much help in ridding our schools of ineffective teachers.

The solution as to the proper applicability of VAM, however, is also not difficult to discern. VAM can continue to serve an essential research purpose, of course, and this role will likely expand as assessments and data capacity improve. In fact, we might view the history of value-added modeling as a story of progress. It begins with researchers responding to the limitations of status-based accountability under the No Child Left Behind Act (NCLB)[5] by advocating for the use of growth models, gain score models, and covariate adjusted regressions, and it progress to the growing use of multilevel, multiyear models such as the layered model and the cross-classified model. These new models allow researchers and evaluators to address old limitations through considering multiple years of data and the cross-nested structure of data. More recently, we have seen the development of variable persistence models that acknowledge and model the research findings that teacher effects partially decay over time. From this perspective, VAM is a story of technical breakthroughs and increasing sophistication. Perhaps, however, we can suggest that there is more work to be done, more advancements that are needed, before VAM can be reliably and validly used as a high-stakes evaluative tool that might dictate the direction or termination of a teacher's career.

The next needed breakthrough might be a model that explicitly acknowledges the latent nature of the traits we are seeking to measure. Student learning and teacher quality are not observable characteristics, but latent constructs. In other words, student learning is not perfectly measured by student test scores. Instead, what we want students to learn encompasses a much broader array of outcomes and cannot be perfectly measured at all. Therefore, researchers should seek to develop structural equation models (SEMs) that take this latent nature into account and explicitly account for measurement error by incorporating various value-added estimates of important student outcomes. We offer one possible structural equation model incorporating value added estimates of higher-order thinking, writing, and emotional health, as well as more typical measures of achievement in Figure 3.

---

[5] NCLB requires that schools be evaluated based on the percentage of students who attained a score of proficient or above on the state standardized tests. Each school is required to make Adequate Yearly Progress (AYP) each year, meaning that each year, each school must raise the percentage of its students who attain proficiency by a set amount. Schools that do not "make AYP" face sanctions including being forced to pay for individual tutoring for any student and being forced to notify parents that they can transfer their students to a "higher-achieving" school. One of the aspects of this law that many experts have criticized is that this system does not take into account the growth students may be making, but only looks at whether students have attained the proficiency bar or not.

In Figure 3, teaching quality is conceived of as a latent (unobserved) trait that produces a causal impact on student learning. Teaching quality is measured through ratings on observations, surveys, and a portfolio (perhaps of classroom artifacts, feedback on student work and lesson and unit plans). These measures are observed, but are conceptualized as being caused by the latent trait of teaching quality, which is why the arrows point from the latent trait to the observed measures. Similarly, student learning is conceived of as a holistic latent trait that is measured through growth on standardized achievement tests, growth on higher-order tests of critical thinking, growth in ratings of student writing, and growth on a survey measuring emotional health. Factor analysis is used to produce an estimate of the latent traits based on the observed measures, and then student learning is regressed on teaching quality to produce an estimate of the effect of teaching quality on student learning.

An SEM, like the one we have proposed in Figure 3, would have the advantage of including multiple student outcome measures, thereby blunting at least one of the major critiques of VAM—narrowing of the curriculum. As opposed to the current VAM-based systems that incentivize a focus on the results of one test, a system that included multiple outcomes could potentially remind us all to conceptualize education holistically and provide incentives to teachers and schools to focus attention on many aspects of student growth.

VAM's utility for evaluative purposes can also be enhanced if we use it as a fire alarm or flag that will focus attention on the "bottom x percent" of teachers. Those teachers could then receive additional outside evaluations by trained experts, and perhaps be required to compile a portfolio of their work, videotape multiple lessons, or submit written reflections on their practice. With such a system, the low value-added estimate would be just a piece of corroborative evidence in the event that a district was seeking to remove a teacher. Also, the additional observations and reflections might help some of these teachers to improve. In all, low performing teachers would be more likely to either improve their practice or leave the classroom without a legal fight if they were confronted with multiple sources of evidence and supported by rich data about their teaching.

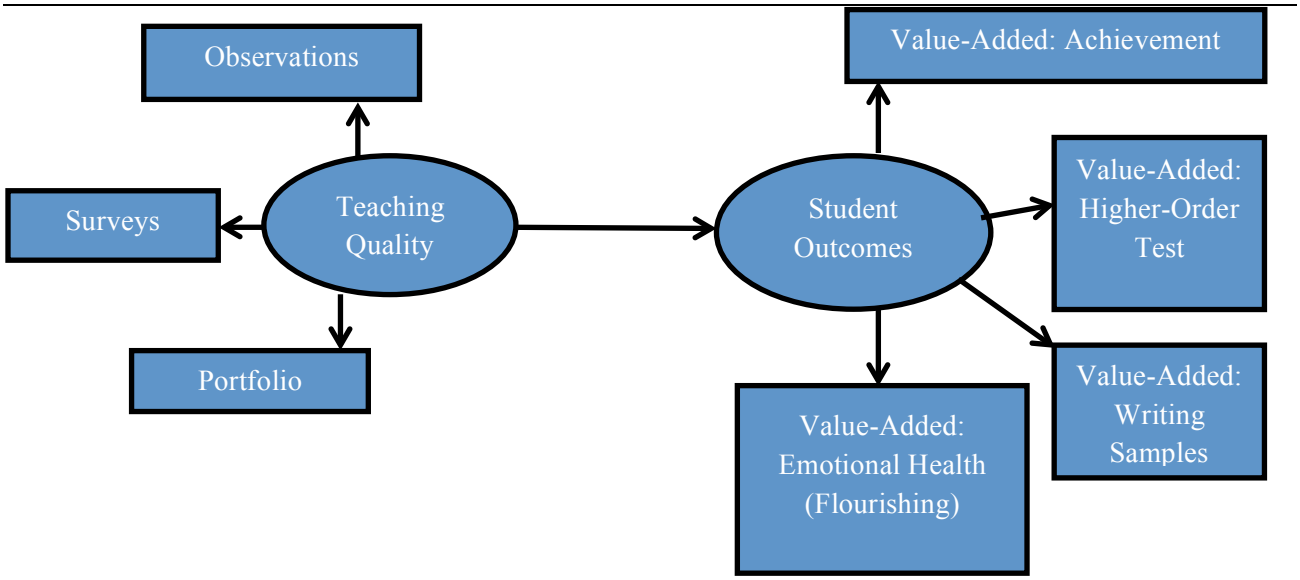

Figure 3. A Possible Structural Equation Model Incorporating Multiple Value-Added Estimates

Such a system could be modeled after the successful Peer Assistance and Review (PAR) programs in Toledo, Rochester, Cincinnati, Columbus, and Montgomery County, or upon the national evaluation system in Chile. One of the major drawbacks to these systems is that they are expensive, but VAM estimates could be used to target extra observations or other evaluative screenings toward those most likely to need assistance or to need to be removed. In such a system, VAM estimates could also serve an important role in helping to support principal ratings. Principals whose ratings consistently diverge from VAM estimates could be accompanied by a second rater to calibrate their scores or given additional training. Finally, randomly selected and perhaps top value-added teachers could also be provided with second raters, thus adding reliability and validity to the overall evaluation system.

Conclusion

VAMs may yet have a bright future in improving our educational systems, but the manner in which they are currently being employed appears to fly in the face of their demonstrated limitations and inconsistencies. While unions and management battle over the percentage that value-added estimates should play in teacher evaluation systems, our findings, and the weight of the research, suggest that a suitable and helpful role for value-added modeling might be better found

outside of such a narrowly-defined summative evaluation system. Value-added modeling can be a tool for researchers to examine the success of policy and curricular changes. Value-added modeling can be a means for calculating growth estimates that could be included in developing a more complex and nuanced picture of student learning, such as might be done using a structural equation model. And value-added modeling can provide estimates that could serve as a fire alarm to alert district supervisors to teachers who might need assistance or additional scrutiny. But, value-added modeling cannot provide us with an incontrovertible answer about how effective any one teacher is. VAM is only a statistical tool, and as such, inferences are dependent on the model used and the assumptions made.

References

Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, 25(1): 95–135.

Baker, E. L., Barton, P. B., Darling-Hammond, L. Haertel, E., Ladd, H. F., Linn, R. L.. . .., Shepard, L. A. (2010, August 29). Problems with the use of student test scores to evaluate teachers. *Economic Policy Institute Briefing Paper*, No. 278. Retrieved from http://s2.epi.org/files/page/-/pdf/bp278.pdf.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011, December). The long-term impacts of teachers: teacher value-added and student outcomes in adulthood. (NBER Working Paper No. 17699). Cambridge, MA: *National Bureau of Economic Research*. Retrieved from http://www.nber.org/papers/w17699.pdf?new_window=1

Corcoran, Sean P. (2010). Can teachers be evaluated by their students' test scores? Should they be? The use of value-added measures of teacher effectiveness in policy and practice. (Education Policy for Action Series). Providence, R.I.: Annenberg Institute for School Reform at Brown University. Retrieved from http://www.scribd.com/doc/37648467/The-Use-of-Value-Added-Measures-of-Teacher-Effectiveness-in-Policy-and-Practice.

Doran, H. C., & Lockwood, J. R. (2006). Fitting value-added models in R. *Journal of Educational and Behavioral Statistics*, 31(2), 205-230.

Felch, J., Song, J., & Smith, D. (2010, August 14). Who's teaching LA's kids? *Los Angeles Times*. Retrieved from http://www.sbcsseport.org/published/a/cp/acpa/collection/29/9/upload.c-acpa-29n9.pdf

Follman, J. (1992). Secondary school students' ratings of teacher effectiveness. *The High School Journal.* 75(3), 168-178.

Green, J. L. (2010). Estimating teacher effects using value-added models. University of Nebraska-Lincoln. Dissertations and Theses in Statistics. Paper 6. Retrieved from http://digitalcommons.unl.edu/statisticsdiss/6

Hanushek, E. A. (2009). Teacher deselection. In Goldhaber, D. and Hannaway, J. (Eds.), *Creating a new teaching profession* (pp. 165–180). Washington, DC: Urban Institute Press.
Retrieved from http://www.latimes.com/media/acrobat/2009-10/49898689.pdf

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *The American Economic Review*, 100(2), 267-271.

Hanushek, E.A. (2011). Valuing teachers: how much is a good teacher worth. *Education Next*, 11(3). Retrieved from http://educationnext.org/valuing-teachers/

Jacob, B. A., Lefgren, L., & Sims, D. P. (2010). The persistence of teacher-induced learning. *Journal of Human Resources*, 45(4), 915-943.

Koedel, Cory and Julian R. Betts (2007). *Re-examining the role of teacher quality in the educational production function*. Working Paper #2007-03. Nashville, TN: National Center on Performance Initiatives. Retrieved from http://economics.missouri.edu/working-papers/2007/wp0708_koedel.pdf

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V. N., & Martinez, J. F. (2007a). The sensitivity of value‐added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*, 44(1), 47-67.

Lockwood, J., McCaffrey, D., Mariano, L., & Setodji, C. (2007b). Bayesian methods for scalable multivariate value-added assessment. *Journal of Educational and Behavioral Statistics*, 32(2), 125–150. Retrieved from http://jeb.sagepub.com/content/32/2/125.full.pdf+html.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin,* 68(5), 304-305.

Lord, F. M. (1969). Statistical adjustments when comparing preexisting groups. *Psychological Bulletin*, 72(5), 336-337.

Mariano, L. T., McCaffrey, D. F., & Lockwood, J. (2010). A model for teacher effects from longitudinal data without assuming vertical scaling. *Journal of Educational and Behavioral Statistics*, 35(3), 253–279.

McCaffrey, D. F., Lockwood J. R., Koretz, D., Louis, T. A. & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, 29(1), 67-101.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating Value-Added Models for Teacher Accountability.* Monograph. RAND Corporation. Retrieved from http://rand.org/content/dam/rand/pubs/monographs/2004/RAND_MG158.pdf

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education,* 4(4), 572-606.

McCaffrey, D.F. (2010). VAM Short Course. Presentation Slides, Data, and Code for R and Stata. Unpublished manuscript, provided by the author.

MET Project (2013). *Ensuring fair and reliable measures of effective teaching: culminating findings. Policy & Practice Brief*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://metproject.org/downloads/MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf

Mihaly, K., McCaffrey, D. F., Staiger, D. O., & Lockwood, J. R. (2013). *A composite estimator of effective teaching*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://www.nbexcellence.org/cms_files/resources/Jan%202013%20%20A%20Composite%20Estimator%20of%20Effective%20Teaching%20Research%20Paper.pdf.

Newton, X., Darling-Hammond, L. Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: an exploration of stability across models and contexts. *Educational Policy Analysis Archives*, 18 (23). Retrieved from http://files.eric.ed.gov/fulltext/EJ913473.pdf.

Papay, J. P. (2011). Different tests, different answers: the stability of teacher value-added estimates across outcome measures. *American Educational Research Journal*, 48(1), 163-193.

Pike, G. R. (2004). Lord's paradox and the assessment of change during college. *Journal of College Student Development*, 45(3). 348-353.

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods (2nd ed)*. Newbury Park, CA: Sage.

Rothstein, J. (2010). Teacher quality in educational production: tracking, decay, and student achievement. *The Quarterly Journal of Economics*. 125(1), 175-214.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, 73(2). 417–58.

Rubin, D. B., Stuart, E. A., & Zanutto. E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics*, 29(1), 103-116.

Sanders, W. L., & Horn, S. P. (1994). The Tennessee value-added assessment system (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education*, 8(3), 299-311.

Sanders, W. L. & Rivers, J.C. (1996). *Cumulative and residual effects of teachers on future academic achievement.* Research Progress Report. Knoxville: University of Tennessee Value-Added Research and Assessment Center.

Sass, T.R. (2008). *The stability of value‐added measures of teacher quality and implications for teacher compensation policy*. CALDER Brief 4. National Center for Analysis of Longitudinal Data in Education Research. Retrieved from http://www.utla.net/system/files/researchbrief2_vam-2009.pdf

Schochet, Peter Z. and Hanley S. Chiang (2010). *Error rates in measuring teacher and school performance based on student test score gains* (NCEE 2010-4004). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://files.eric.ed.gov/fulltext/ED511026.pdf

Tekwe, C. D., Carter, R. L., Algina, J., Ma, C. X., Lucas, M. E., Roth, J., Ariet, M., Fisher, T., Resnick, M. B. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics*, 29(1), 11-36.

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: *The New Teacher Project*. Retrieved from http://widgeteffect.org/downloads/TheWidgetEffect.pdf.

Wiley, E.W. (2006). A practitioner's guide to value added assessment. Educational Policy Studies Laboratory Research Monograph. Tempe, AZ: Arizona State University. Retrieved from http://nepc.colorado.edu/files/Wiley_APractitionersGuide.pdf.

Authors

Kevin Schaaf is a writer, researcher, and Ph. D. student in education, focusing on social research methodology. He is the father of four children, taught elementary school for 9 years, and holds a Master's in Education and a Master's in Public Policy. His research interests center on improving teaching, including his work on mathematics instruction and formative assessment at the National Center for Evaluation, Standards, and Student Testing (CRESST), past research on professional development, and this issue's focus on teacher evaluation.

Daniel Dockterman is a third year Ph.D. student studying Social Research Methodology in the School of Education at the University of California, Los Angeles. He is an Advanced Quantitative Methods fellow and currently works at the National Center for Research on Evaluation, Standards and Student Testing (CRESST) evaluating the IMPACT urban teacher residency program. Daniel's research interests include multilevel modeling and measures of teacher quality, specifically value-added modeling and student and teacher surveys.