

Value-Added Models: Analytic Issues

A paper prepared for the National Research Council and the National Academy of Education, Board on Testing and Accountability Workshop on Value-Added Modeling, November 13 and 14, 2008, Washington D.C.

Daniel F. McCaffrey and J.R. Lockwood

The RAND Corporation

The term value-added modeling (VAM) comes from the economic literature on the estimation of the contribution of current inputs to outputs after accounting for prior inputs. In the education literature, VAM refers to efforts to measure the effects on student achievement of students' current teachers and schools separate from other inputs. More broadly, the phrase is often used to describe any analysis using longitudinal student test score data to study the effects of educational inputs on achievement. The term is also often associated specifically with the analysis of sources of variation in student growth in achievement, although as demonstrated below, the current VAM approaches go far beyond straightforward analysis of simple gain scores (current year less prior year test scores).

While the origins of VAM of teacher effects date back over 30 years (Hanushek, 1972; Murnane, 1975), interest in these methods among researchers, policy makers, and educators grew precipitously following the publication of a technical report by William Sanders and June Rivers in 1996 that found teacher effects estimated using student test score gains predict student outcomes at least two years into the future, suggesting that teachers have persistent effects on their students' achievement and the accumulation of these effects could be substantial. The following year, Sanders and his colleagues published another paper claiming that teachers are the most important source of variance in student achievement. Interest in VAM was further stoked by replication of the Sanders and Rivers results (Mendro et al., 1998; Rivers, 1999) and several other papers finding that the variability among teachers was large and that value-added estimates of teacher effects predict teachers' future students' outcomes (Aronson, Barrow and Sander, 2007; Gordon, Kane and Staiger, 2006; Harris and Sass, 2005; Lockwood et al., 2007a; Nye, Konstantopoulos, and Hedges, 2004; Rivkin, Hanushek and Kain, 2005; Rowan, Correnti and Miller, 2002). The growing availability of student achievement data due to increased emphasis on test-based accountability and the passage of the No Child Left Behind Act in 2001 led to explosive growth in empirical research on the effects of teachers on student outcomes and value-added methods in the last five years.

Value-added models have been proposed for four main purposes: (1) school and teacher improvement, (2) school and teacher accountability, (3) program evaluation, and (4) research. Harris (forthcoming) combined the first two purposes into what he calls value-added modeling for accountability (VAM-A) and examples include programs in Ohio, Tennessee, Dallas and New York City that estimate individual teacher effects for diagnostic purposes and increased use of these estimates for high-stakes decisions such as teacher pay, including programs in Nashville, Florida, Houston, and other locations. Harris calls the latter activities (program evaluation and research) value-added modeling for program evaluation (VAM-P) and examples include assessment of classroom practices in mathematics and science and studies on teacher attributes such as National Board of Professional Teacher Standards certification (National Research Council, 2008).

Analytic Challenges for VAM

Although often tacit, the goal of value-added modeling is to make what quantitative researchers refer to as *causal inferences* or to estimate *causal effects*. The causal effect measures the change in outcomes caused by an intervention or other agent such as a teacher or school. In VAM, the goal is to determine how students' learning and achievement differ having been in their assigned school or teacher's classroom or participated in a program rather than being taught in an alternative school, by an alternative teacher or not participating in the program. Although

causal estimation is a common goal in social science, what sets VAM apart from many other applications is the use of longitudinal student achievement data to make its causal estimates. VAM purports to overcome the challenges of causal estimation by using repeated measures of student achievement to account for student, family and prior educational inputs into education.

Causal estimation requires the estimation of counterfactual quantities such as the expected outcomes for students taught by teacher A had they been taught by teacher B and vice versa. The challenge for estimating causal effects via observational data is ensuring that groups receiving different educational inputs provide estimates of these counterfactual quantities that are not confounded by pre-existing differences between groups. This challenge is particularly pronounced for VAM because different schools and classrooms often serve very different student populations and programs are typically targeted at particular groups of students. The challenge for VAM is then to remove these differences via use of multiple achievement tests. In the language of Campbell and his colleagues (Shadish, Cook, and Campbell, 2002), VAM must ensure the internal validity of the quasi-experiment by using longitudinal data.

The challenges of causal estimation are even greater for VAM-A applications because educators and the public are highly aware of the differences among schools in the student populations in terms of family and neighborhood backgrounds and that these inputs are strong correlates of student achievement and account for much of the variability in student achievement. VAM-A estimates will be heavily scrutinized for evidence that they do indeed level the playing field for schools and teachers, and any evidence that they fail to do so could erode their acceptance by educators, destroy their face-validity, and limit their utility for motivating improvements in education via accountability or performance-based pay.

Beyond the challenges of removing potential confounds to ensure internal validity, VAM also faces the challenge of construct validity. Educational inputs are generally conflated so that a classroom of students might receive inputs from the school, the principal, the teacher, other teachers in the school, the community, other students within the classroom, and potentially other sources. Similarly, teachers participating in different interventions or programs or with different credentials and qualifications tend to differ on multiple dimensions that might influence student learning. Teasing apart this multitude of inputs is likely to be impossible (Raudenbush and Wilms, 1995). Hence even if estimates are made from classrooms or schools serving identical populations, the interpretation of the resulting causal effects is likely to be difficult. In fact, even defining causal effects of interest can be challenging: what teacher effects should be estimated if teacher effectiveness is determined in part by school leadership?

A related challenge for estimating value-added for schools is determining whether the causal effect of interest is annual or cumulative and how to estimate cumulative effects when students might not have data prior to entering the school, as is the case in elementary schools.

For VAM to be useful, the estimates must not only be unbiased but also precise. In VAM-A, applications estimates from year to year must also be reasonably stable. Instability will erode face-validity because most researchers and education practitioners will expect that true teacher performance will change gradually over time rather than display huge swings from year to year. Moreover, if estimates are unstable, they cannot be used to guide or motivate appropriate changes in future behavior through evaluation or compensation incentives.

A final challenge for VAM-A is to balance the need for analytic complexity to achieve accurate estimation against the reality that a goal of accountability is to provide educators with signals about what is considered good performance and whether they have achieved such performance, as well as to motivate lower-performing individuals to change their behavior to

improve their effectiveness. There is general agreement that highly complex statistical procedures are difficult for educators to understand and concern that the use of such procedures to achieve accurate estimates might limit the utility of VAM. Accepting less accuracy for more transparency in the estimation methods might actually improve the functioning of VAM-based accountability or compensation systems, but how to achieve the correct balance remains unknown.

Analytic Approaches to VAM

Current approaches to VAM grew out of three distinct traditions: statistics, economics, and *ad hoc* methods.

Ad hoc Methods

Ad hoc methods are typically developed by individual schools, school districts, or states for their particular needs for VAM-A. These measures might share features with methods from the other traditions but they also might include nonstandard analytic procedures or combine standard measures with other empirical evaluations. For example, one school district is using average percentage gains in achievement to evaluate school performance for compensating principals because the measure places greater value on lower-performing students. Another district is measuring school performance by combining value-added measures with the percentage of students who graduate from high school.

The use of *ad hoc* methods appears to be motivated by the desire to provide measures that offer strong signals to teachers about expected performance. Educators may also accept *ad hoc* measures because they appear trustworthy and transparent (i.e., they seem to measure what they are supposed to measure). The statistical properties of these measures are rarely considered and may be extremely difficult to assess.

Statistical Methods

Statistical models provide a description of repeated measures of student achievement that captures the important statistical features of the data such as the average growth across students, variance around the average, and correlation among scores that share common features such as scores from the same students or scores from different students who share or have shared a classroom or school.

The models describe the correlation among students who share or shared common classrooms by assuming that achievement test scores can be decomposed into additive random components associated with each classroom or school (current or prior) and residual terms that depend on the student and the year of testing. For example, letting A_{it} denote a student's achievement in a given subject area (e.g., mathematics) at time t , a simple representation of a statistical model is

$$A_{it} = m_t + q_{it} + q_{it-1} + \dots + q_{it1} + e_t \quad (1)$$

where m_t equals the mean or average for all students at this time, q_{it} equals a random variable that is common to all students who share the student's classroom at time t , q_{it-1} equals a random variable that common to all students who shared the student's classroom at time $t-1$, etc., and e_t denotes a residual error term describing how the student's score at time t deviates from the

average of all other students and all other students who share his or her history of classroom assignments. It also describes how the student's achievement at time t deviates from his or her achievement in other years.

Analysts have taken multiple approaches to modeling the components for prior year classroom or school membership (q_{t-1} to q_{t1} in model (1)). Some analysts (Sanders, Saxton, and Horn, 1997, Raudenbush and Bryk, 2002) assume these components do not change over time so that the component for fourth grade classroom membership on fourth grade scores is that same as it is on fifth, sixth, seventh, and all future grade scores. The fourth grade classroom leaves an indelible mark on student achievement test scores that persists unchanged through the remaining years of testing. McCaffrey et al. (2004) and Lockwood et al. (2007a) allow for the effect to diminish over time by a constant that depends on the year of testing and the year of the classroom or school membership but which does not differ across students. Recently, Mariano, McCaffrey, and Lockwood (2008), extended this model further to allow the components to be distinct but correlated random variables.

A key feature of the statistical models is the correlation of the residual error terms, e_i in model (1), from repeated measures on the same student. Some analysts (Sanders, Saxton, and Horn, 1997, McCaffrey et al., 2004, Lockwood and McCaffrey, 2007) make no assumptions about the correlation across years, e.g., a student's seventh and eighth grade scores are correlated but the model does not assume a particular structure for the correlation. Other researchers (Raudenbush and Bryk, 2002) assume more structure among the residual errors by describing their growth as a direct function of time (e.g., achievement grows linearly with time and the intercept and slope are specific to the student and vary randomly across students). Regardless of the assumptions for the structure of the correlation, the correlation among the error terms is the feature of the statistical model that leads to an adjustment for students' prior achievement when estimating the current year school, teacher or classroom effect.

Statistical Models and Causal Effects

An important characteristic of the statistical models is that they are descriptive. They are not structural and make no explicit attempt to relate to counterfactuals needed for causal inferences. The components for classrooms or schools are random variables meant to capture commonalities among scores for students who shared a school or classroom. Many sources can contribute to these components, including not only teachers or schools, but also all the other schooling and *non-schooling* factors that might be unique to this classroom (school) and common among the students. However, even though the models are not explicitly causal, using them to assess teachers and schools implicitly treats them as causal effects.

The descriptive models will be causal models under unrealistic assumptions that students are effectively randomly assigned to classrooms (schools), so that none of the commonality in scores from students who share classrooms (schools) is due to disproportionate allocation of students with certain characteristics to different classrooms (schools). This assumption is clearly not true in the vast majority of schools and school districts in our country and hence it cannot be used to support a causal interpretation of the estimates from the statistical models.

Lockwood and McCaffrey (2007), however, establish more realistic conditions under which the estimates of the classroom components from these descriptive statistical models will recover nearly unbiased estimates of the causal effects of the classroom. Roughly speaking, the conditions require a large number of tests and that student-level factors that are associated with

classroom assignment can be fully described by low-dimensional latent variables. For instance, suppose that students have a latent general level of achievement, i.e., some students generally score high every year and other students score lower every year. Furthermore, suppose that classroom assignment depends on this general level of achievement but not on other student-level factors, then the assumptions of Lockwood and McCaffrey (2007) would be met. If, on the other hand, classroom assignment depends on this general level of achievement and other factors such as performance on any particular test or classroom behavior during the prior school year, then the assumptions of Lockwood and McCaffrey (2007) would not be met. Through simulations, Lockwood and McCaffrey (2007) show that with as few as roughly 5 test scores, the statistical models could provide estimates of causal effects that have very limited bias due to student assignments. Their results also apply to VAM-P applications provided that at least some individual students are assigned to the program of interest in some years and the control in others; this switching of “treatment” status over time is also the essential identification strategy of standard econometric approaches to VAM-P using longitudinal achievement data.

Even if the assumptions of Lockwood and McCaffrey (2007) are met, the causal effects of the classroom (school) assignment are not necessarily a causal effect of the teacher. As discussed previously, many factors vary by classroom and which ones contribute to student learning remains unknown. Some of these factors might be related to the context of the neighborhood and local involvement in the school and classroom. Others could be other schooling factors or complex interactions between the teacher and the context that may or may not be related to the causal effect of the teacher or the causal effect of interest. Statistical models make no efforts to untangle these varied inputs and effects can have a particular causal interpretation only under the assumption that other inputs have at most very weak effects.

The properties of the estimates of statistical models have generally been assessed by simulation studies, like those described above, in which estimates from a statistical model are compared to the true values under varying assumptions about the data and their relationship to a causal model. Another approach to evaluating the methods has been to study estimated effects from alternative models to see how they compare and how they relate to student background characteristics aggregated to the classroom level.

These studies have tended to find mixed results. As noted above, Lockwood and McCaffrey (2007) created some scenarios where the estimates recovered the causal effects in their simulated data. However, McCaffrey et al. (2004) created scenarios where this did not occur. In the latter scenarios, classrooms were stratified such that there were disjoint subpopulations of classrooms and students in which none of the students in one stratum ever shared a classroom with any of the students in other strata. In this situation the statistical models cannot recover causal effects if student growth differs on average by stratum. Recent work by McCaffrey, Sass, and Lockwood (2008) showed that economic models are also potentially biased by stratification but stratification appeared to be a very limited problem in large urban school districts in Florida, so this source of bias might be inconsequential in practice.

Authors have found that estimated classroom effects (or what they call “teacher effects”) are sensitive to the specification of prior year classroom components in achievement scores from subsequent years of testing. For example, the correlation of estimates from models that assume perfect persistence of these components and those that allow them to degrade is about .8 in empirical comparisons made by Lockwood et al. (2007a).

In addition, studies have found somewhat mixed results from relating estimates to aggregated student characteristics. Sanders (2006) reports weak correlation with demographics.

McCaffrey, Han and Lockwood (2008) also found weak correlation, but comparisons across many alternative VAM procedures suggest that estimates of classroom (or teacher) value-added from the statistical models estimated with data on Nashville middle school mathematics teachers might be favoring teachers of classes with students who tended to be higher-achieving prior to entering the teachers' classrooms.

The statistical models that have been used in practice have tended not to include student-level covariates such as race or socio-economic status measures. One argument for excluding covariates from models is that including them might imply different expectations for students of different socio-demographic or other groups. However, there are also technical challenges to including such variables in the model. The decision to exclude them from the models has been criticized by researchers and practitioners, which resulted in an empirical investigation of the effects of including these variables in the models. That study (Ballou, Sanders, and Wright, , 2004) developed a method for including student-level covariates in the models that avoided the technical problems and found that their inclusion had no appreciable effect on estimates of classroom effects. Attempts to expand the methods to include classroom-level variables resulted in unstable estimates (Ballou, 2005). In general, statistical models cannot yield causal estimates that remove the effects of classroom-level variables.

Economic Models

The econometric modeling approach to VAM emerged out of the cumulative achievement model. The models start with a general education production function for student achievement:

$$A_{it} = A_i[\mathbf{F}_i(t), \mathbf{E}_i(t), \mu_{i0}, \varepsilon_{it}] \quad (2)$$

where A_{it} , the achievement level for individual i at the end of his or her t th year of schooling, depends on the cumulative achievement function A_i and inputs: the entire input histories of family (including neighborhood) and school-based educational inputs $\mathbf{F}_i(t)$ and $\mathbf{E}_i(t)$, respectively, a composite variable representing time-invariant characteristics an individual is endowed with at birth (such as innate ability), μ_{i0} , and a normally distributed, mean-zero error, ε_{it} .

Economists make a series of assumptions to specify the production function model (2) in a form that supports estimation of teacher effects using established estimators. In particular, they assume that: the cumulative achievement function A_i does not depend on age and is additively separable; family inputs are time-invariant and can be combined into a single term with student inputs; educational inputs can be separated into additive annual contributions of schools, teachers, peers and other classroom inputs. Without further assumptions, estimation of the value-added of the current teacher would require data on the school, teacher, peer, and classroom inputs for each student's current and entire prior schooling experience. Because these data do not exist, economists typically assume that all educational inputs decay geometrically at a rate d and that family inputs change at that same rate. These assumptions lead to the model:

$$A_{it} = g_s \mathbf{S}_{it} + g_T \mathbf{T}_{it} + g_p \mathbf{P}_{it} + g_c \mathbf{Z}_{it} + j_i + d A_{it-1} + \varepsilon_{it}, \quad (3)$$

where \mathbf{T}_{it} denotes teacher inputs and can be a single factor for a teacher effect, \mathbf{S}_{it} , \mathbf{P}_{it} , \mathbf{Z}_{it} denote

school, peer and other classroom inputs, \mathbf{j}_i is the fixed contribution of student and their family, and the error term is $\mathbf{x}_{it} = \varepsilon_{it} - I\varepsilon_{it-1}$. This fixed student-family contribution includes all the unobservable characteristics of the student and family (including community context) that contribute to achievement and are stable across time. A final assumption that $I = 1$ leads to circumstances where a linear model for the first differences in achievement (gain scores) yields consistent estimators of certain effects, including those of the current teachers. The model parameters are estimated using least squares with indicators for students and teacher (i.e., student and teacher fixed effects) and other variables in the model. Todd and Wolpin (2003) and Harris and Sass (2005) provide detail derivation of this model and the assumptions required to achieve model (3).

There are applications that use this model for both VAM-A and VAM-P. Because I is unknown, applications commonly assume $I = 1$; however, some applications have used $I = 0$ which then results in fixed effects analysis on level scores rather than gain scores. In other applications, I is estimated by including the lagged scores as a predictor in a linear regression model. Some of these applications have treated \mathbf{j}_i as zero and some have attempted to estimate both I and the student coefficients while accounting for the complexities this creates for estimation (Koedel and Betts, 2007).

Model (3) lacks specificity about the other schooling inputs that are needed to fully specify the structural model. Given that administrative data typically have limited measures of specific schooling inputs, it can be challenging to include all the necessary measures in the model when estimating teacher effects and this misspecification could bias the estimated effects. One approach sometimes used to avoid under-specifying school inputs is to include school indicators in the estimation and include school means in the structural model. However, this creates an estimation problem since the average teacher mean at the school is now confounded with the school mean. Computationally this is solved in estimation by defining teacher causal effects as relative to the average effect for the school and estimating these effects, by including both school and teacher indicator variables into the linear models used for estimation.

Although such a strategy allows for estimation of a well defined causal effect, making this change is not without consequences. Teachers would now be evaluated by direct comparison with their colleagues in their school. However, this type of direct comparison of teachers in the same school is generally considered unacceptable for accountability and generally the school fixed effects cannot be used VAM-A (Harris, 2007). Rather, models using available data on school inputs are used with known potential for bias from misspecification. Using school fixed effects might be appropriate for VAM-P and could improve estimation, provided that programs are not school-level.

Economic Models and Causal Effects

Unlike statistical models, which are descriptive and without explicit links to structural models or causal effects, the economic models are developed as models for the outcomes of all students had they received the educational inputs given in \mathbf{S}_{it} , \mathbf{P}_{it} , and \mathbf{Z}_{it} . The economists identify assumptions under which the parameters of model (3) can be estimated unbiasedly (for large samples) using standard least squares or related approaches. Unbiased causal effect estimates can be obtained directly from the estimated model parameters provided the necessary assumptions all hold.

Of course the fly in the ointment for the economic models is whether or not these assumptions, many of which lack face-validity, actually hold in practice. The economic tradition is to develop empirical tests of the assumptions. The primary challenge of VAM is to remove the potential confounding of estimates due to differences among students receiving different educational inputs and provide internally valid causal effect estimates. In the economic modeling, the critical assumption for ensuring this goal is met is that any differences among classes, schools, or programs that are not captured by variables explicitly used in modeling, are captured by factors included in the static fixed student-family components, β_j s. In the terminology of the economics literature, any selection of inputs on the basis of unobservables is restricted to the static unobservables in the fixed student-family component. Given that this is the key assumption for meeting the primary goals of the model, this assumption has received the most scrutiny among economists. Economists have also explored the rate of decay (λ) and the sensitivity of models to assumptions about this rate and the assumption that inputs have equal effects on all students.

Recent work by Rothstein (forthcoming, 2008) offers the most thorough test of the assumption that assignments to inputs rely only on static unobserved quantities. Rothstein tested this assumption in the context of estimating teacher effects by testing if teachers can predict the achievement gains of their students in the years *prior* to these students being in their classes. For instance, does a fifth grade teacher predict her students' achievement gains when those students were third and fourth graders? He found that teachers did indeed predict their students' prior achievement gains. Since teachers cannot rewrite the past, the finding that teachers predict their students' prior performance implies there is selection of students into teachers' classrooms that is related to student prior achievement growth. Rothstein also found that the relationship between current teachers and prior gains differs across time, i.e., the grade 5 teacher's relationship with grade 4 gains differs from the relationship with grade 3 gains, and he interprets this as evidence that class assignments cannot depend on a single static factor, but rather must be dynamic. This means that at least in some locations, the central assumption of economic modeling does not hold and VAM estimates are likely to be biased. The size of the bias and the prevalence of the conditions leading to the violations are unknown. Although this result was aimed at testing the specification of economic models, it has important implications for the interpretation of estimates from statistical models because dynamic classroom assignment would also violate the assumptions that Lockwood and McCaffrey (2007) establish for allowing causal interpretation of statistical model estimates.

Economic tests of assumptions of the decay find that it is generally large, maybe as much as 50 percent in a year (Kane and Staiger, 2008, Jacob, Lefgren, and Sims, 2008), which is quite contrary to the often-assumed zero decay. However, limited exploration suggests that estimates can be reasonably robust to violations of assumptions about decay. Finally, regarding the assumption that inputs have equal effects on all students, Koedel and Betts (2007) provide some evidence supporting assumptions that teacher effects are reasonably constant across all students and Lockwood and McCaffrey (2008), using a very different modeling technique, also find that any existing interactions appear to be small.

Precision and Stability of VAM Estimates

Research on the precision and stability of estimates is mostly related to VAM-A applications – VAM-P programs and policies generally involve large proportions of teachers or

large administrative databases, so precision is not an issue.

Studies of the precision of VAM-A estimates consistently find that sampling errors in the estimates are large. For example, typically the standard errors are sufficiently large that about two-thirds of estimated teacher effects are not significantly different from average or zero (McCaffrey et al., 2005). Confidence intervals are wide and rankings would likely contain substantial errors, so that only about 30 to 35 percent of teachers ranked in either the top or bottom quintile in one year remain so in the next year (Aaronson et al. 2004; Ballou, 2005). Similarly, studies consistently find that the correlation of estimated effects from pairs of adjacent years are low, ranging from less than 0.2 to typically no higher than 0.5 (Aaronson et al. 2007; McCaffrey, Sass, and Lockwood, 2008)

Variability in estimated teacher effects is not only a result of sampling error due to the modest numbers of students in classes. For instance, McCaffrey, Sass, and Lockwood (2008) decomposed the variability in repeated measures of teacher effects for samples of mathematics teachers from four counties in Florida and found that there was year-to-year variability in teacher effects that exceeded expectations due to simple sampling error from random samples of students in classrooms. This additional source of error generally accounted for a much larger share of the variability in effects for elementary than middle school teachers and this source of variability was only weakly related to administrative data on teachers such as credentials, tenure, and annual levels of professional development. Whether this variability reflects variation in true performance or a source of error at the classroom level remains unknown. However, it does have implications for combining data across years because it suggests that such pooling might mask true variation in performance, especially for elementary teachers.

Scaling of Achievement Measures

All VAM estimates of teacher effects depend on achievement test scores, but different approaches place more or less stringent requirements on these tests. In all cases, value-added estimates can only exist in grades and subject areas in which testing occurs and prior achievement data exist. Also, teachers can only be evaluated on the constructs measured by the achievement tests. Critics sometimes fault the standardized tests which are commonly the foundation of VAM for being too narrow and not measuring higher-order thinking skills and thus question the utility of VAM estimates derived from these test scores.

In addition to potential limitations in the range of content on standardized tests, psychometricians have raised concerns about comparability of scores from different grade-level tests, even when tests are vertically linked to a single scale. The primary concern is that tests at each grade-level measure multiple constructs and shifts in the mix of constructs across grades can distort test score gains, invalidate assumptions of perfect persistence of teacher effects and the use of gain scores to measure growth, and bias VAM estimates.

Some statistical models make less stringent assumptions about scaling than models that assume complete persistence or directly model growth. However, the problems of scaling suggest that current and prior year scores might not even be linearly related because of differential scale compression at the top and bottom of the scale. Hence, assumptions of normality and implicit linearity in the relationships between scores from pairs of years might also be inconsistent with test scores and estimates from statistical models that rely on joint normality might be biased by inappropriate assumptions. Errors about linearity in the relationships among scores from multiple years can be particularly problematic when using these scores to control for

differences among widely disparate classes as is done implicitly in the statistical approaches to VAM. Linearity of scores can be explored empirically and models or scales could be adjusted if the model does not conform to the data.

Common Ground in Analytic Methods for VAM

Currently there is limited consensus in the VAM literature about the best analytic approaches. Advocates of *ad hoc* methods argue that transparency is the primary concern for providing useful VAM-A estimates. However, quantitative researchers from both the statistical and economic modeling traditions argue that complex methods are likely to be necessary for accurate estimation of teacher effects and that accountability or compensation systems based on performance measures with weak statistical properties will fail to provide educators with useful information to guide their practice and could eventually erode their confidence in such systems.

However, quantitative researchers do not agree on the best approach to complex modeling. Economists focus on and develop economic models to explicitly deal with the possible unobservable factors related to assignment of students to classes within a structural model framework, even as there is growing evidence that the other assumptions that those models require about persistence of the effects of historical inputs and scaling of tests may not be met. Statisticians and quantitative education researchers continue to use descriptive statistical models, adding more and more flexible methods for modeling the persistence of effects and relaxing assumptions about scaling issues. However, these augmentations do not directly address the primary concern of VAM that estimates not be confounded by student background variables and that the estimates have construct validity and not conflate many different inputs into teacher or school effects. The shortcomings of each class of models are so potentially problematic that there is no consensus on the best approaches, and little work has been done on synthesizing the best aspects of each approach.

For VAM-P applications, the paradigms are more similar, with both groups relying on measured covariates to control in part for group differences. However, economic applications tend to favor models of gain scores and the use of fixed effects, whereas analysts from the statistical persuasion tend to favor random effects models (e.g., hierarchical linear models) and do not directly account for selection beyond the inclusion of observed variables.

Analysts in both paradigms have been fairly shaken by Rothstein's (2008) results. Currently there has not been a clear response from either paradigm and both camps may need to adapt their modeling approaches to address the possible fallacy of their current modeling assumptions. Both camps also tend to agree that the precision of estimated effects is generally low and that steps need to be taken to improve precision if the estimates are to be used in practice. Currently both camps tend to rely on shrinkage estimation (Morris, 1983), which is implicit in the statistical approach and applied *post hoc* to economic estimates, and combining data across years.

Areas for Future Research

The primary need for future analytic research on VAM is for expanding both the current economic and statistical models to incorporate features from the other paradigm that are missing in their own approaches. For instance, economic models need to be expanded to relax assumptions about constant geometric decay across all inputs and the rate of this decay.

Relaxing these assumptions will also allow the models to make more realistic assumptions about test scaling. Statistical models need to be expanded to explicitly model the assignment of students to classrooms and schools—for instance by allowing the selection to depend on latent random student effects. The models might also need to relax assumptions about the joint normality of vectors of scores.

Both paradigms need to evaluate the implications of Rothstein's results and develop models that can either allow for dynamic assignment on unobserved variables or minimize the bias if removing it proves impossible. Alternatively, analysts might identify additional data that would be required to allow for unbiased estimation and develop instruments and plans for collecting that data as part of general education practice.

There is also a need for continued research on methods to improve the precision of estimated effects. In particular, results suggesting that there is true variability in teacher performance across years suggests that simply pooling data across years might introduce bias and methods that smooth across time while allowing for true deviation in performance (e.g., state-space models (Judge et al., 1985) might be appropriate. Another potentially fruitful approach might be to share information across teachers by including teacher characteristics in the models and shrinking estimates toward the performance predicted by these factors. For example, the model might include years of experience, so that shrinking is toward the average of teachers with similar experience rather than the average for all teachers.

Finally, there is also a significant need to study how teachers use information from compensation and accountability systems to determine the appropriate balance between the complex estimation methods necessary for accurate measures and the need for measures to be transparent. Accuracy is necessary for a compensation system to reward effective teachers and provide estimates that predict future performance to guide under-performing teachers toward effective changes. Transparency enables teachers to understand how performance measures use data and how they relate to their teaching, so that teachers accept such data as valid measures of their performance which respond to factors under their control. Without transparency, teachers may not accept a performance-based compensation system, and the system might have little effect on teachers' practice regardless of the accuracy of the measures. Currently very little is known about how teachers interpret the results of performance measures or performance-based pay. Similarly there is no explicit research about the level of transparency required for a measure to be acceptable to teachers.

Discussion

Although the desired uses for value-added models clearly require causal estimates of teacher effects for a broad population of students, there appears to be more evidence which questions such causal interpretations than evidence which supports them.

Models from the statistical paradigm are descriptive rather than causal, and causal interpretations are somewhat of a leap of faith. Lockwood and McCaffrey (2007) have established conditions under which that leap is justified and have shown via simulation that even with a modest number of tests per student it is possible that bias can be small provided class assignments depend on a few stable attributes of students but not on characteristics that vary from year to year. The work of Rothstein suggests that at least in North Carolina, classroom assignments appear to be dynamic and inconsistent with the conditions Lockwood and McCaffrey (2007) require. Moreover, even if the statistical models do provide causal effects,

those effects conflate teacher inputs with other schooling and non-schooling inputs that affect classrooms.

The economic paradigm specifically builds structural models that describe potential achievement under counterfactual classroom assignments and can provide estimates of causal effects. However, the assumptions required to allow for unbiased estimation of those model parameters are elaborate, often lack face-validity and are not generally well supported in empirical investigations.

However, the research on teacher value-added is far from universally bad. There are many studies that support the use of value-added for assessing teachers. Kane and Staiger (2008) find value-added estimates and estimates of teacher performance on randomly assigned classrooms provide similar evaluations of teachers. Cantrell et al (2007) also find that value-added scores are the best predictors of teachers' future students' outcomes even among pairs of teachers on randomly assigned classes. This result echoes earlier findings by Sanders and colleagues showing value-added measures are correlated with students' future outcomes. Many studies show estimated teacher effects are often not correlated with aggregate classroom demographics and a limited number of studies have found them to be related to principal evaluations, teacher observation protocols, and tests of teacher knowledge. Moreover, the implications of violations of assumptions in terms of how large the bias may be and the prevalence of the conditions that lead to violations is not known and may not be severe.

But the case for teacher value-added may be more about what it provides than whether or not it provides it perfectly. Teachers matter and they are not all equally effective. The traditional credential and qualification proxies for teacher effectiveness do not appear to provide good evidence about teachers' ability to produce student learning. The current evaluation system based on observations has been widely criticized as ineffective and unable to provide sound guidance about how to improve teaching practices, or which teachers to tenure into the profession (Wilson, forthcoming; Danielson and McGreal, 2000). Newer observation-based protocols rely on ideas of "good teaching" based in educational theory with only limited evidence that these measures can truly identify teachers who most (or least) effectively promote student learning.

Given the weaknesses in the current evaluations, measures of teacher effects on student outcomes could benefit the teaching profession by guiding effective teachers to continue best practices and directing others to seek professional development. Accurate measures of teacher effects on student learning could also serve as the basis for compensation and accountability systems which might motivate teachers who are struggling to produce student learning gains to seek the advancement opportunities they require, reward the most effective teachers in response to younger teachers' and young potential teachers' demands for being rewarded on the basis of their merits, and attract a broader range of professionals to the profession.

Other uses of student test scores suffer from all the same shortcomings of value-added and more. Thus, even though value-added is potentially less than ideal, it may have significant advantages over other test-based accountability and evaluation systems.

However, because of the mixed results on the accuracy of value-added measures, a universal rollout of systems using value-added measures for high-stakes decisions such as compensation and tenure is probably not well supported by the current research, unless those programs include well developed evaluation plan and plans to stop such pilots if they appear harmful to teachers or students. VAM-based systems need to be introduced into educational practice in careful and thoughtful ways. Educators need to be trained in how to use value-added

information and its limitations. Programs using it need to be piloted and evaluated through randomized experiments in which teachers whose performance is measured by value-added are randomly assigned to the intervention while their peers are assigned to control conditions. An example of such a pilot program is the POINT experiment being carried out by the National Center on Performance Incentives in Nashville to study performance pay.

Other types of evaluation are also possible. For example, shadow tenure decisions could be made in a value-added system used by researchers to identify teachers who would or would not be tenured if value-added were used in the decision process. Assuming that all these teachers would most likely get tenure under the systems currently in place (since nearly all teachers are indeed tenured), all teachers could be tracked as they advanced in their careers and the performance on value-added and other measures of teachers designated for tenure could be compared to those who were not.

Finally, there is a need for continued development of analytic methods that rely on fewer assumptions for estimating unbiased effects and a need for greater data collection so that data rather than assumptions can be used to identify teacher or school effects. More studies like the inventive experiment of Kane and Staiger would also be valuable. All this research will be supported by greater use of value-added in practice, and so, regardless of their many limitations, it seems that greater use of value-added measures rather than less is the only way forward.

References

- Aaronson, D., Barrow, L., and Sander, W. (2007). Teachers and student achievement in the Chicago Public High Schools. *Journal of Labor Economics*, 25, 95–135.
- Ballou, D., (2005). Value-added assessment: Lessons from Tennessee. In R. Lissetz (Ed.), *Value Added Models in Education: Theory and Applications*. Maple Grove, MN: JAM Press.
- Ballou, D., Sanders, W., and Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, 29(1), 37–65.
- Cantrell, S., Fullerton, J., Kane, T. J., and Staiger, D. O. (2007). *National board certification and teacher effectiveness: Evidence from a random assignment experiment*. Unpublished paper. Available: <http://harrisschool.uchicago.edu/Programs/beyond/workshops/ppepapers/fall07-kane.pdf> [accessed October 16, 2008]
- Danielson C., and McGreal, T. L. (2000). *Teacher evaluation to enhance professional practice*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Gordon, R., Kane, T. J., and Staiger, D. O. (2006). *Identifying effective teachers using performance on the job* (Discussion Paper 2006–01). Washington, DC: Brookings Institution.
- Hanushek, E. (1972). *Education and Race*. Lexington, MA: D.C. Heath and Company.
- Harris, D.N. (forthcoming-). The policy uses and “policy validity” of value-added and other teacher quality measures. In D. H. Gitomer (Ed.), *Measurement issues and the assessment for teacher quality*. Thousand Oaks, CA: SAGE Publications.
- Harris, D. (2007). Diminishing marginal returns and the production of education: An international analysis. *Education Economics*, 15(1), 31-45.
- Harris, D. and Sass, T. (2005). Value-added models and the measurement of teacher quality. A paper presented at the 2005 conference of the American Education Finance Association.
- Jacob, B. A., Lefgren, L., and Sims, D. (2008). The persistence of teacher-induced learning gains. Unpublished manuscript.
- Judge, G.G, Griffiths, W. E., Hill, R.C., Lütkepohl, H. and Lee, T.-C. (1985). *The theory and practice of Econometric, Appendix C. The Kalman Filter*. New York: John Wiley and Sons.
- Kane, T.J. & Staiger, D.O. (2008). *Are teacher-level value-added estimates biased? An experimental validation of non-experimental estimates*. A paper presented at the National Conference on Value-Added Modeling, Madison, WI, April 22-24, 2008.

- Koedel, C., and Betts, J.R. (2007). *Re-examining the role of teacher quality in the educational production function*. Working Paper #2007-03. Nashville, TN: National Center on Performance Initiatives.
- Lockwood, J.R., and McCaffrey, D.F. (2007). Controlling for individual heterogeneity in longitudinal models, with applications to student achievement. *Electronic Journal of Statistics*, 1, 223-252. Available at: <http://dx.doi.org/10.1214/07-EJS057>.
- Lockwood, J.R., and McCaffrey, D.F. (2007). Are teachers differentially effective with students of differing abilities? A paper presented at the National Conference on Value-Added Modeling, Madison, WI, April 22-24, 2008.
- Lockwood, J.R., McCaffrey, D.F., Mariano, L.T., and Setodji, C. (2007a). Bayesian methods for scalable multivariate value-added assessment. *Journal of Educational and Behavioral Statistics*, 32(2), 125-150.
- Mariano, L.T., McCaffrey, D.F., and Lockwood, J.R. (2008). A model for teacher effects from longitudinal data without assuming vertical scaling. Submitted, *Journal of Educational and Behavioral Statistics*.
- McCaffrey, D.F., Han, B., and Lockwood, J.R. (2008). From Data to Bonuses: A Case Study of the Issues Related to Awarding Teachers Pay on the Basis of Their Students' Progress. A paper presented at Performance Incentives: Their Growing Impact on American K-12 Education, February 28-29, 2008, Vanderbilt University Nashville, TN.
- McCaffrey, D.F., Lockwood, J.R., Koretz, D., Louis, T., and 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., Mariano, L. T., and Setodji, C. (2005). Challenges for value added assessment of teacher effects. In R. Lissitz (Ed.), *Value added models in education: Theory and practice* (pp. 272–297). Maple Grove, MN: JAM Press.
- McCaffrey, D.F., Sass, T.R., and Lockwood, J.R. (2008). The intertemporal effect estimates. A paper presented at the National Conference on Value-Added Modeling, Madison, WI, April 22-24, 2008.
- Mendro, R.L., Jordan, H.R., Gomez, E., Anderson, M.C., and Bembry, K.L (1998). *An application of multiple linear regression in determining longitudinal teacher effectiveness*. Paper presented at the 1998 Annual Meeting of the AERA, San Diego, CA.
- Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association*, 78, 47-55,

- Murnane, R.J. (1975). *The Impact of School Resources on the Learning of Inner City Children*. Cambridge, MA: Ballinger Publishing Co.
- National Research Council. (2008). *Assessing Accomplished Teaching: Advanced Level Certification Programs*. Committee on Evaluation of Teacher Certification by the National Board for Professional Teaching Standards. Milton D. Hakel, Judith Anderson Koenig, and Stuart W. Elliott, editors. Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.
- Nye, B., Konstantopoulos, S., and Hedges, L.V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, 26(3), 237-257.
- Raudenbush, S.W., and Bryk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd Edition). Thousand Oaks, CA: Sage Publications.
- Raudenbush, S.W., and Wilms, J.D. (1995). The estimation of school effects *Journal of Educational and Behavioral Statistics*, 20, 307-335.
- Rivers, J.C. *The Impact of Teacher Effect on Student Math Competency Achievement*, dissertation, The University of Tennessee, Knoxville, 1999, Ann Arbor, Mich.: University Microfilms International, 9959317, 2000.
- Rivkin, S. G., Hanushek, E., and Kain, J. F. (2005). Teachers, schools and academic achievement. *Econometrica*, 73, 417–458.
- Rothstein, J. (forthcoming). Student sorting and bias in value added estimation: Selection on observables and unobservables. *Education Finance and Policy*.
- Rothstein, J. (2008). Teacher quality in educational Production: Tracking, Decay, and Student Achievement,” working paper.
- Rowan, B., Correnti, R., and Miller, R.J. (2002). What large-scale survey research tells us about teacher effects on student achievement: Insights from the Prospects study of elementary schools. *Teachers College Record*, 104, 1525-1567.
- Sanders, W. L. (2006). Comparisons among various educational assessment value-added models. A paper presented at The Power of Two--National Value-Added Conference, Battelle for Kids, Columbus, Ohio, October 16, 2006.
- Sanders, W.L., and Rivers, J.C. (1996). *Cumulative and residual effects of teachers on future academic achievement*, Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center, November 1996.
- Sanders, W., Saxton, A., and Horn, B. (1997). The Tennessee Value-Added Assessment System: A quantitative outcomes-based approach to educational assessment, In J. Millman (Ed.),

Grading teachers, grading schools: Is student achievement a valid evaluation measure? (pp. 137-162). Thousand Oaks, CA: Corwin Press, Inc.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin Company.

Todd, P. E., and Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *Economic Journal*, 113, F3–F33.

Wilson, S. (forthcoming). Measuring teacher quality for professional entry. In D. H. Gitomer (Ed.), *Measurement issues and assessment for teaching quality*. Thousand Oaks, CA: Sage Publications.