Consortium for Educational Research and Evaluation– North Carolina

Comparing Value-Added Models for Estimating Individual Teacher Effects on a Statewide Basis

Simulations and Empirical Analyses

Roderick A. Rose, Department of Public Policy & School of Social Work, University of North Carolina at Chapel Hill

Gary T. Henry, Department of Public Policy & Carolina Institute for Public Policy, University of North Carolina at Chapel Hill

Douglas L. Lauen, Department of Public Policy & Carolina Institute for Public Policy, University of North Carolina at Chapel Hill

August 2012

Table of Contents

COMPARING VALUE-ADDED MODELS FOR ESTIMATING INDIVIDUAL TEACHER EFFECTS ON A STATEWIDE BASIS: SIMULATIONS AND EMPIRICAL ANALYSES

Executive Summary

Many states are currently adopting value-added models for use in formal evaluations of teachers. We evaluated nine commonly used teacher value-added models on four criteria using both actual and simulated data. For the simulated data, we tested model performance under two violations of the potential outcomes model: settings in which the single unit treatment value assumption was violated, and settings in which the ignorability of assignment to treatment assumption was violated. The performance of all models suffered when the assumptions were violated, suggesting that none of the models performed sufficiently well to be considered for high stakes purposes. Patterns of relative performance emerged, however, which we argue is sufficient support for using four value-added models for low stakes purposes: the three-level hierarchical linear model with one year of pretest scores, the three-level hierarchical linear model with two years of pretest scores, the Educational Value-Added Assessment System (EVAAS) univariate response model, and the student fixed effects model.

Introduction

A wide body of research into the effects of schooling on student learning suggests that teachers are the most important inputs and, consequently, that improving the effectiveness of teachers is a legitimate and important policy target to increase student achievement (Rockoff, 2004; Nye, Konstantopolous, & Hedges, 2004; Rowan, Correnti, & Miller, 2002). In order for education policymakers and administrators to use teacher effectiveness to achieve student performance goals, they must have accurate information about the effectiveness of individual teachers. A relatively recent but often recommended approach for obtaining teacher effectiveness estimates for use in large-scale teacher evaluation systems relies on *value-added models* (VAMs) to estimate the contribution of individual teachers to student learning; that is, to estimate the amount of gains to student achievement that each teacher contributes rather than focusing on levels of student achievement (Tekwe, Carter, Ma, Algina, Lucas, et al., 2004). These VAMs rely on relatively complex statistical methods to estimate the teachers' incremental contributions to student achievement. Value-added models could be viewed as primarily descriptive measurement models or putatively causal models that attribute a portion of student achievement growth to teachers (Rubin, Stuart, & Zanutto, 2004); we take the latter view in this study.

Proponents maintain that VAM techniques evaluate teachers in a more objective manner than by observational criteria alone (Harris, 2009). By holding teachers to standards using *outcomes*, policymakers could move away from standards based on inputs in the form of educational and credentialing requirements and principals' or others' more subjective observations of teachers' practices (Gordon, Kane, & Staiger, 2006; Harris, 2009). There are concerns that VAMs may not be fair appraisals of teachers' effectiveness because they may attribute confounding factors, unrelated to instruction, to the teacher (Hill, 2009). Further, evidence suggests that teacher effectiveness scores may vary considerably from year to year (Sass, 2008; Koedel & Betts, 2011), despite teachers' contentions that they do not vary their teaching style (Amrein-Beardsley, 2008), suggesting that the year-to-year variability is unrelated to teacher effectiveness. While the controversies about the accuracy and utility of VAMs continue to swirl, many states have agreed to incorporate measures of teacher effectiveness in raising student test scores into their teacher evaluations in order to receive federal Race to the Top (RttT) funds or achieve other policy objectives. The uses of teacher VAM estimates in the evaluation process vary from low stakes consequences, by which we mean an action such as developing a professional development plan; to middle stakes, by which we mean actions such as identifying teachers for observation, coaching, and review; to high stakes, by which we mean denial of tenure or identifying highly effective teachers for substantial performance bonuses. In spite of the commitment by many states to use a VAM for estimating teachers' effectiveness, there is no consensus within the research community on the approach or approaches that are most appropriate for use. Given these concerns and the widespread use of these models for teacher evaluation, evidence on the relative merits of VAMs is needed.

Several techniques for estimating VAMs have been compared using simulated or actual data (Guarino, Reckase, & Wooldridge, 2012; Schochet & Chiang, 2010; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Tekwe et al., 2004). Tekwe et al. (2004) used actual data, while McCaffrey et al. (2004) used both simulation data and actual data. Simulation studies have used either correlated fixed effects (Guarino, Reckase, & Wooldridge, 2012; McCaffrey et al.,

2004) or variance decomposition frameworks for data generation (Schochet & Chiang, 2010). To date, no study has used both correlated fixed effects and variance decomposition simulated data as well as actual data. The present study aims to provide a more comprehensive assessment and uses all three types of data. Moreover, the present study compares nine common VAMs, more than in any other study published to date. Finally, we compare VAMs using the rankings of teachers in the true and estimated distributions of effectiveness using four distinct criteria that are relevant to policymakers, administrators, and teachers.

We compare these nine VAMs using simulated and actual data based on criteria that include their ability to recover true effects, consistency, and false positives in identifying particularly ineffective teachers (or ineffective teachers; the results are nearly identical). To determine which VAMs best handle the confounding influence of non-instructional factors and identify ineffective teachers, we generate simulation data, with known teacher effectiveness scores to compare with teacher effectiveness estimates from each VAM. We use actual data from a statewide database of student and teacher administrative records to examine the consistency between models and relative performance of each VAM in year-to-year consistency in teacher ranking.

In this study, we take the view that teacher effect estimates from VAMs are putatively causal, even though the conditions for identifying causal effects may not be present. Therefore, we first discuss the potential outcomes model (Reardon & Raudenbush, 2009; Rubin, Stuart, & Zanutto, 2004; Holland, 1986). Subsequently, we introduce seven common VAMs and then review existing studies comparing VAMs for teacher effect estimation in the context of the potentially unrealistic demands placed on these VAMs by the potential outcomes model. We then discuss the methods in the present study, including the data generation process for both simulations and the characteristics of the actual data, the form of the nine models compared, and the comparison techniques. We follow with the results of these comparisons. In the final section, we discuss the implications of these findings for additional research into VAMs for teacher effect estimation and implementation as a teacher evaluation tool.

The Potential Outcomes Model

Value-added models, in economic terms, measure the output resulting from combining inputs with technology (i.e., a process; Todd & Wolpin, 2003). If estimates from VAMs of student assessment data are to be inferred as and labeled *teacher effect estimates* then they should be viewed as causal estimates of teachers' contributions to student learning. That is, the value added estimands are not simply descriptions of students' average improvement in performance under a given teacher, but are effect estimates causally attributed to the teacher. This view coincides with the use of VAMs in education policies such as teacher evaluation. It is widely acknowledged that the process by which the teacher causes student learning does not have to be specified (see, for example, Todd & Wolpin, 2003; Hanushek, 1986). It is not as widely understood that the process by which students learn does not have to be fully specified in order to identify a causal teacher effect. The causality of the estimand from a VAM can instead be derived from assumptions that are independent of model specification (Rubin, Stuart, & Zanutto, 2004).

The assumptions of the potential outcomes model, if met, support the causal inference of teacher effect estimates from VAMs (Reardon & Raudenbush, 2009; Rubin, Stuart, & Zanutto, 2004; Holland, 1986). The central feature of the potential outcomes model is the counterfactual —the definition of the causal estimand of a teacher's effect on a student depends on what the student experiences in the absence of the specified cause—that is, under any other teacher besides the one to which the student was assigned. This enables us to ignore inputs to cumulative student knowledge that are equalized over different treatment conditions and are not confounded with treatment assignment. A formal model for causality begins as follows. First, assume that the outcome for student i (with i = 1,…N) under teacher j is Y_{ij}. Second, assume that each j^{th} teacher is a separate treatment condition from J possible treatments, and each student has one potential or latent outcome Y_{ij} under each possible teacher (of which at most one can actually be realized). This is a many-valued treatment (Morgan & Winship, 2007) with the potential outcomes represented by a matrix of N students by J treatments (Reardon & Raudenbush, 2009). Because only one such treatment can be identified (the fundamental problem of causal inference; Holland, 1986), the treatment effect is defined as a function of the distributions of students assigned to teacher j and the students under any other teacher. Generally, this is implemented using linear models based on the average treatment effect for teacher j (ATE_j) comprised of students observed under assignment to teacher j compared to the other teachers, which we label as $.j$ (not j), e.g., a simple mean difference $d_{ij} = \Delta_{ij} - \Delta_{i.j} = E[Y_{ij}] - E[Y_{i.j}]$. An obvious candidate for $.j$ is the teacher at the average level of effectiveness.

Reardon and Raudenbush (2009) identified six defining, estimating, and identifying assumptions of causal school effects that they suggested are also appropriate for teacher effects, two of which we make explicit here. Defining assumptions include (1) each student has a potential outcome under each teacher in the population (manipulability); and (2) the potential outcome under each teacher is independent of the assignment of other participants (the stable unit treatment value assumption, or SUTVA). Estimating assumptions include (3) students' test scores are on an interval scaled metric; and (4) causal effects are homogeneous. Identifying assumptions, when satisfied, make it possible to infer the treatment effect as causal despite the fundamental problem of causal inference that only one of J potential outcomes can be realized. These assumptions

include (5) strongly ignorable or unconfounded assignment to teachers; and (6) each teacher is assigned a "common support" of students, which may be relaxed to assume that each teacher is assigned a representatively heterogeneous group of students, to estimate an effect that applies to all types of students. This last assumption may alternatively be met by extrapolation of any teacher's unrepresentative group of students to students that the teacher was not assigned if the functional form of the model (e.g., linear or minimal deviations from linearity) supports such extrapolation.

Building on the formal model of causality discussed above, this section presents a formal discussion of two of the six assumptions of the potential outcomes model that are relevant to the comparison between VAMs in the present study, drawing heavily on Reardon and Raudenbush (2009).

Stable Unit Treatment Value Assumption (SUTVA)

SUTVA implies that the treatment effect of any teacher on any student does not vary according to the composition of that teacher's classroom (Rubin et al., 2004):

(i) $Y_{ij}(AZ) = Y_{ij}$

AZ is an N x J matrix of ij elements recording the assignment of students to teachers, with $ij = 1$ if i is assigned to j and $ij = 0$ otherwise. The statement above makes it explicit that Y_{ij} is invariant to all Z permutations of A, a vector indicating each student's assignment to treatment. Ruled out by this assumption are effects based on composition of the classroom, including those attributable to peer interactions and those between peers and teachers. Therefore, a student assigned to a math classroom with higher achieving peers should have the same potential outcome under that teacher's treatment as they would if the classroom contained lower achieving peers. The effects that classroom composition may have on learning make this assumption challenging to support. For example, if teachers alter instruction based on the average achievement level of the class, these effects imply that the treatment effect for a single student is heterogeneous according to the assignment of peers (Rubin et al., 2004).

Ignorability

The second assumption, ignorability, implies that each student's assignment to treatment—that is, their assignment to a specific teacher (A)—is independent of their potential outcome under that teacher (Morgan & Winship, 2007):

(ii) $Y_{ij} \perp A$

The formal statement ii is the strictest version of this assumption, strong ignorability. A weaker form conditions on student background factors, x:

(ii') $Y_{ij} \perp A \mid X = x$

Because each VAM handles the specification of these covariates differently, the weaker form of the assumption will affect the comparability of the models with the true teacher effect. The

strong form is violated under any associations with measured or unmeasured covariates; the weak form is violated because of the association between omitted variables—unmeasured inputs to learning—with (1) each student's potential outcomes, and (2) each student's assignment to teachers (Reardon & Raudenbush, 2009). This is a common problem because large-scale administrative data files do not contain measures of all of the inputs to learning (including such predispositions as motivation, for example, or the availability of reading material in the home) and because students are not usually randomly assigned to teachers. These problems manifest as associations between the average characteristics of students assigned to a teacher and the teacher effect. The ignorability assumption makes it explicit that unmeasured inputs to student learning are not confounders of the teacher effect unless these unmeasured inputs are also associated with treatment (student-teacher) assignment. While still a demanding assumption, this nevertheless substantially reduces the burden for satisfying causality.

Violations of Assumptions

Non-ignorable assignment of students to teachers is widely viewed as the key problem that VAMs must address. The most widely accepted form of ignorable assignment is randomization. The quantitative evidence suggests that observed student and teacher characteristics are associated, and thus regardless of what the assignment process is (such as a tracking mechanism or a compensatory assignment policy), it may not be effective at equalizing over students' propensity to perform on end-of-grade tests or teachers' propensity to promote student achievement (Rothstein, 2010; Clotfelter, Ladd, & Vigdor, 2006). That is, it is not equivalent to a random process. An alternative to randomization is to adjust the models for the factors that are associated with both assignment and student outcomes, an approach that is dependent upon model specification. The information needed to sufficiently adjust the value-added models for these confounding effects in the presence of non-random assignment may not be known or available in the form of quantified measures. Longitudinal data files spanning multiple years on both the students and teachers, and linking students to teachers, are required in order to estimate these models. In many states these datasets are available, but certain information, like students' home and family background characteristics, may not be measured, though proxies for these characteristics (e.g., subsidized lunch for family income) may be.

Interactions among peers in attenuating or magnifying a teacher's effectiveness, a manifestation of the violation of SUTVA, may also confound estimation of a teacher's true level of effectiveness. As a consequence of this violation, the risks are high that the estimate of the teacher effect is actually some combination of the teacher effect and these peer learning effects. The problems manifesting under violations of this assumption should not be confused with those related to assignment. SUTVA violations occur after assignment and regardless of the satisfaction of the ignorability of assignment. Just as in the case of ignorable assignment, the data typically used to estimate teacher effects do not contain the information needed to directly measure these interactions. Instead, the data may contain only proxies for these interactions, such as the average prior performance of peers in a class. Proxies for manifestations of SUTVA violations, like average performance, may also be evidence of assignment non-randomness.

Under recent federal mandates, such as eligibility for Race to the Top funding, state education agencies may be required to use VAMs to rank and compare teachers across an entire state (Henry, Kershaw, Zulli, & Smith, 2012). In many cases, these agencies may also impose

requirements on principals that they incorporate VAM estimates into their teacher evaluation criteria with high-stakes consequences for teachers. These consequences may include being placed under a performance review or being dismissed. Regardless of the particular model selected for generating value-added teacher estimates, therefore, the intended uses of these models imply that VAM estimates are causal effects of teachers on student achievement growth. Data limitations and the failure of assumptions for causality to hold in practice raise important questions about teacher effect estimation and whether the estimates can in fact be interpreted as causal effects for which teachers should be held accountable. Most empirical analyses, including VAMs, are based on theoretical assumptions such as those of the potential outcome model that are rarely met in practice. The effects of these deviations from theoretical assumptions should be studied to determine how they affect conclusions about teacher performance. Further, with several models available from the literature on value-added modeling, comparisons between models on the differential effects of deviations should be undertaken. The present study, like several other studies that precede it, represents such an attempt. Before discussing the studies that this effort is intended to build upon, we formally present and describe seven typical value-added model specifications.

Typical Value-Added Models

In order to facilitate a discussion of the existing body of research comparing value-added models with each other or with some absolute standard, we present seven value-added models from a broad selection of statistical and econometric specifications that are most typically found in the literature. We provide the model specification, define the teacher effect in each model, and state the practical implications within each specification of satisfying the key assumptions from the potential outcome framework. For consistency, we use common notation for each model, despite variations in the source material. There are three large classes of VAMs being widely used: (1) nested random effects models, (2) econometric fixed effects models, and (3) hybrid fixed and random effects models, such as the educational value-added assessment system (EVAAS; Ballou, Sanders, & Wright, 2004). (It is important to bear in mind that "fixed effects" refers to using unit-specific dummies or demeaning to control for unmeasured confounders and not to the coefficients on student covariates.) We use this organizing framework in the sections that follow.

Nested Random Effects Models

Nested random effects models treat the multiple levels of data—student, classroom, teacher, and school—as hierarchically nested units, where each lower level (e.g., student) can only be nested within one higher level unit. As such, these are generally single-year cross-sectional models, though the effects of previous years' teachers on individual students' current performance are typically accounted for by pretest covariates that are included in the model. These include hierarchical linear models and multilevel models (Raudenbush & Bryk, 2002). These models are based on the following general specification:

$$(1) \qquad Y_{ijsw} = X_{ijs}\beta_{js} + X_s\beta_s + \beta_{w-1}Y_{i,w-1} + \beta_{w-2}Y_{i,w-2} + u_s + u_{js} + e_{ijs}$$

Subscripts indicate the student (i), teacher (j), school (s), and period (w). The variable Y is a test score on a standardized end-of-grade exam in a selected subject area; X_{ijs} is a vector of time-invariant student characteristics or predispositions to learning that are associated with the accumulation of knowledge, as well as a constant term; X_s is a vector of school characteristics; w = the period for which the teacher effect is being estimated; $w-1$ = the prior period; $w-2$ = two periods' prior. Therefore, Y appears both as the dependent variable in the current period, as well as predictors of student achievement in the current period as both a one-year and two-year lag or pretest. The u and e terms are errors for (in order) the school, the teacher, and the student. The teacher effect is estimated from the empirical Bayes residual or empirical best linear unbiased predictor estimate of u_{js} (the teacher random effect). Variations on this model exist, including those that ignore the nesting of students and teachers within schools or use fewer prior test scores. All of the models of this type use a random effects model to estimate the teacher effect and may use covariates at the student and school levels to control for correlates of non-random assignment. To satisfy the ignorability requirement of the potential outcomes model, u_{js} cannot be associated with excluded covariates that are also associated with the outcome. Because u_j are errors, all factors associated with both assignment and the outcome must be measured and included in the model. This is a strong and potentially impractical requirement.

Fixed Effects Models

The second major type of VAM consists of a variety of fixed effects specifications. Fixed effects are within-group variance estimators frequently used in econometric models in which students or teachers (or both) are used to constrain variance in the context of panel data. The ignorability assumption is generally satisfied by way of indirect controls for confoundedness (or in econometric terms, endogeneity) rather than via specification of the confounding factors in the model. Student fixed effects models use students as their own control and aggregate within-student variation only. Teacher fixed effects models are similar, but rather than use teachers as their own controls, each teacher effect is explicitly estimated. In both cases, the models are assumed to be adjusted for confounders that do not vary "within" (over time for student fixed effects, or over students for teacher fixed effects). Alternative specifications, developed from Arellano and Bond (1991) and used in previous VAM research (Guarino et al., 2012), add instrumental variables (IV) in the form of twice-lagged measures of the dependent variable (the outcome two periods removed) to circumvent endogeneity between time-varying inputs, manifesting on the one-year lagged outcome, and the teacher effect. There are four major variations on fixed effects models: (1) a student fixed effects model; (2) a teacher fixed effects model; (3) a student fixed effects instrumental variable model; and (4) a teacher fixed effects instrumental variable model. Numerous models appear throughout the econometric evaluation literature but are largely variations on these major types.

The student fixed effect model (SFE) uses a multi-year panel with demeaning of all characteristics:

$$(2) \qquad (Y_{iw} - \bar{Y}_i) = (\mu_{iw} - \bar{\mu}_i) + (\alpha_i - \bar{\alpha}_i) + (e_{iw} - \bar{e}_i)$$

The terms with bars (e.g., $\bar{Y}_i$) are the within-student means of each parameter. The term μ_{iw} represents time-varying predictors of student achievement; the student fixed effect α_i, which absorbs the fixed effects of time-invariant unmeasured confounders, such as predispositions to learning, is accordingly eliminated by demeaning. Alternatively, time-varying effects of these predispositions, if they exist, are not. The teacher effect is estimated as the mean of the composite residuals within each teacher, $(e_{iw} - \bar{e}_i)$. To satisfy the ignorability requirement of the potential outcomes model, this complex error must not be associated with any of the terms in the model, including serially.

The teacher fixed effects model (TFE) is a cross-sectional model much like the random effects models, focusing on the nesting of students in teachers but estimated using teacher dummy variables. Unlike the SFE, the teacher effect parameters in this case are estimated:

$$(3) \qquad Y_{ij} = X_{ij}\beta_j + \beta Y_{i,w-1} + \alpha_j + e_{ij}$$

The teacher fixed effect model bears a resemblance to some nested random effects models; instead of estimating the teacher effect with the random effect u_j, the teacher effect is estimated using the dummy variables represented by α_j, and the difference between the random and fixed effects estimates of the teacher effect is due to Bayesian shrinkage in accordance with the reliability of each teacher's sample average (Raudenbush & Willms, 1995). To satisfy the

ignorability requirement of the potential outcomes model, α_j must not be associated with excluded covariates that are also associated with the outcome.

Student and teacher fixed effects control for, respectively, between-student variation and between-teacher variation. The student fixed effects model does not address within-student variation, which is controlled for via covariates. The potential outcomes model informs us that within-student factors that are associated with assignment to treatment as well as the outcome represent confounders that are therefore not accounted for without explicitly including them in the model. If unobserved, they manifest on the error. The same applies to the teacher fixed effects model, though in that case, it is between-student variation at any point in time that represents the source of confounding.

Other econometric models use both fixed effects and instrumental variables in a two-stage least squares framework in order to eliminate endogeneity not eliminated by the fixed effects (i.e., serial or time varying effects in the student fixed effects model, such as the student's previous year test score). Rather than demeaning over the entire panel, the fixed effect component is estimated using a "first difference." With the two-stage framework, a first difference of the lag or pretest period is estimated with the outcome two periods removed (twice lagged) as the instrument. Subsequently, the predicted value of this first differenced pretest is entered as a covariate into the model for the current period, with first differencing:

$$\Delta Y_{i,w-1} = \Delta X_i\beta + Y_{i,w-2} + \alpha_j + r_i \quad (4.1)$$

$$\Delta Y_i = \Delta X_i\beta + \Delta\hat{Y}_{i,w-1} + \alpha_j + e_i \quad (4.2)$$

First differencing, much like demeaning, eliminates the fixed effects of time invariant characteristics. In the student fixed effect IV model (SFEIV), the term α_j is not directly estimated; the teacher effects are calculated from the mean of the residuals within each teacher. In the teacher fixed effect IV (TFEIV), the teacher effect is estimated as the term α_j. As in the previous specifications of the SFE, the SFE teacher effects, derived from the residual, must not be associated with any other terms in the model. For the SFEIV, e_{ij} must not be associated with α_j, contemporaneously. In both cases, serial correlation in e_i is addressed by the instrument. Note also that the IV specification assumes, in contrast to the random effects models and other fixed effects models, that the outcome two periods removed satisfies the exclusion restriction: that its effect on the outcome operates strictly through the endogenous variable and does not have a net direct effect on the outcome used to estimate the teacher effect. This implies that the relationship between 3rd grade performance and 5th grade performance must be completely mediated by 4th grade performance.

A final model is a simple pooled regression model (labeled by Guarino, Reckase, & Wooldridge [2012] as a dynamic ordinary least squares, or DOLS) that uses the panel of data but ignores the nesting of time within students, treating each observation as independent:

$$Y_{ijw} = X_{ijw}\beta_j + \beta Y_{i,w-1} + \alpha_{jw} + e_{ijw} \quad (5)$$

The DOLS bears a resemblance to both the HLM2 and TFE models, but uses the panel of data over multiple years instead of treating each grade level as a separate cross section. The teacher effects are estimated for all grade levels from the α_{jw} term. To satisfy the ignorability requirement of the potential outcomes model, the teacher effect α_{jw} must not be associated with excluded covariates that are also associated with the outcome.

Hybrid Fixed and Random Effects Models

The hybrid approach uses both random effects to estimate the teacher effect as an empirical Bayes residual or shrinkage estimate (Wright, White, Sanders, & Rivers, 2010) as well as fixed effects—either unit-specific dummy variables or demeaning, though usually the former—to control for confounders of the teacher effect. Covariates are generally not entered into the models (Ballou, Sanders, & Wright, 2004).

The multivariate response model (MRM) is also a type of multiple membership, multiple classification (MMMC) random effects model (Browne, Goldstein, & Rasbash, 2001). MMMC or cross-classified models are multilevel or random effects models that acknowledge the complex nature of student-teacher assignment over multiple years of schooling or multiple subjects (Browne, Goldstein, & Rasbash, 2001). That is, rather than being simply nested within a single teacher, students are nested within multiple teachers over time, but the students with which they share this nesting changes. The MRM can be represented in a very simplified form for a student in three sequential grade levels (3rd through 5th) as follows (Ballou, Sanders, & Wright, 2004):

$$(6.1) \quad Y_3 = b_3 + u_3 + e_3$$

$$(6.2) \quad Y_4 = b_4 + u_3 + u_4 + e_4$$

$$(6.3) \quad Y_5 = b_5 + u_3 + u_4 + u_5 + e_5$$

Here, 3rd grade cumulative learning for this student (Y_3) is viewed as the baseline year with an average test score of $b_3$within (for example) the state, a 3rd grade teacher contribution (u_3), and a 3rd grade random error (e_3). The teacher contribution and error are therefore deviations from the state average. Fourth grade cumulative learning is decomposed into a state average (b_4) and deviations from this average in the form of both 3rd and 4th grade teacher contributions (u_3 and u_4) and the 4th grade random error. Fifth grade cumulative learning is similarly decomposed into the state average (b_5), teacher effects in grades 3, 4, and 5, and an error in grade 5 (e_5). Each teacher's effect therefore is assumed to persist without decay into the future, and the MRM is frequently referred to as a layered model because of the appearance that teacher effects in later periods are layered on top of those from earlier periods. More advanced versions of the MRM (e.g., Wright, White, Sanders, & Rivers, 2010) incorporate partial teacher effects for settings where a student had more than one teacher for a subject. In addition, they may also include fixed effects dummy variables for subject, year, and grade level. The u terms, the teacher effects, are estimated as empirical Bayes residuals. These residuals satisfy the ignorability assumption if they are not associated with unmeasured correlates of the outcome. The layering in the MRM

requires making very strong assumptions about teacher effect persistence; teacher effects cannot be attenuated by later teachers or decay of their own accord, and they cannot increase over time.

The MRM has limited practical application on a statewide level because of the high computational resources it requires (McCaffrey, et al., 2004). Consequently, its application has been more limited, being conducted on a districtwide basis. Alternatively, for statewide level models, an alternative model, the univariate response model (URM), has been used. The URM is a version of the EVAAS hybrid model that accommodates measuring student growth using tests that are not developmentally scaled, such as high school end-of-course tests (Wright, White, Sanders, & Rivers, 2010). Like the MRM, covariates other than prior test scores in multiple subjects are not included in the model. Fixed effects dummy variables are not incorporated into the model; instead, fixed effects are incorporated via de-meaning of lagged scores or pretests using a multi-step process. First, the difference between each student score at time *w*, *w-1,* and *w-2*, and the grand mean of school means at each time point is calculated. The exams from which these scores are obtained do not need to be on the same developmental scale as the outcome, or as each other. Second, the covariance matrix C, here shown as being partitioned into current test score (y) and lagged test score (x) sections, is estimated:

$$(7.1) \quad C = \begin{bmatrix} \mathrm{C_{yy}} & \mathrm{C_{yx}} \\ \mathrm{C_{xy}} & \mathrm{C_{xx}} \end{bmatrix}$$

The expectation-maximization algorithm is used to estimate C in the presence of conditionally random missing values. Third, the coefficients of a projection equation b are estimated as follows:

$$(7.2) \quad b = C_{xx}^{-1}\mathrm{c_{xy}}$$

Fourth, the following projection equation is estimated using the elements of b as the $\hat{\beta}$, which predicts a composite of students' previous test scores, spanning two years and two subjects, that have been recalibrated using de-meaning or mean-centering as pooled-within-teacher values (m = math and r = reading):

$$(7.3) \quad C_i = \hat{\mu}_y + \hat{\beta}_{m1}(x_{im1} - \hat{\mu}_{m1}) + \hat{\beta}_{m2}(x_{im2} - \hat{\mu}_{m2}) + \hat{\beta}_{r1}(x_{ir1} - \hat{\mu}_{r1}) + \hat{\beta}_{r2}(x_{ir2} - \hat{\mu}_{r2})$$

The x_{ikt} terms (with $k = m$ for math and $k = r$ for reading; and $t = 1$ for a one-period lag and $w = 2$ for a two-period lag) are each student's test scores in the specified periods and subjects. The $\hat{\mu}_{vw}$ terms are means of the teacher means of these test scores, with $\hat{\mu}_y$ as the mean in the current period, and $\hat{\beta}_{m1}$ are the elements of b (7.2). Finally, substitute the composite C_i into the following two-level model (students nested in teachers), and just as in the previous multilevel models, estimate the teacher effect using the empirical Bayes residual:

$$(7.4) \quad y_{ij} = \beta_0 + \beta_1 C_{ij} + u_j + e_{ij}$$

The nesting in this final model is of students within teachers in one school year with no accounting for the nesting within schools. In addition, the teacher effect estimation uses only one subject, despite the use of two subjects' data to estimate the composite C_{ij}. To satisfy the

ignorability requirement of the potential outcomes model, u_j must not be associated with unmeasured factors that are also associated with the outcome. This means that C must subsume all such factors. This is a strong and potentially impractical requirement. As far as we know, the URM, which is the VAM currently used by EVAAS on a statewide basis, has not been evaluated in any peer-reviewed study.

Summary of Models

All of the models except for two (the TFE and DOLS) obtain the teacher effect indirectly via a type of residual, either a parameter variance (in the random coefficients models) or error variance (in the fixed effects models). The random effects models, as well as the teacher fixed effect model, must control for student unobserved heterogeneity through covariates, available or measured inputs to learning or proxies for inputs to learning that may be highly correlated with the actual inputs. The student fixed effects model, which controls for heterogeneity by demeaning such that time-invariant confounding factors are eliminated, does not address confounding time-varying factors. The IVE models add an additional control for confounding time-dependent inputs by instrumenting a twice-lagged test score. None of the models explicitly addresses SUTVA, though some of the models may make accommodations for violations of these assumptions. For example, the random effects models could include an additional parameter variance for the classroom level, which would distinguish between effects due to the teacher and effects due to the teacher's interaction with each group of pupils to which they are assigned in each year, thereby directly addressing the SUTVA assumption. The teacher fixed effects model could also include interactions between each teacher indicator and student background characteristics, although the model would get even more unwieldy than it already is with J teacher effects being estimated.

Many of these models for estimating teacher effects have been subjected to a rigorous comparison for estimating teacher effects, while others have not. Ideally, these comparisons would be made with regard to an absolute standard, the true teacher effect. Because the true rankings are not typically known, scholars have relied largely on simulation studies in which true effects are generated by a known process and then compared to the estimated effects obtained from VAMs on these data. Actual data have been used, alternatively, to assess relative performance, including consistency across years in the ranking of each teachers, and to assess the plausibility of the ignorability assumption. We now review these studies.

VAM Comparison Studies

Several studies have compared the performance of different methods—random and fixed effects methods, for example—for estimating VAMs. Some of these studies have used actual data, while others have used simulated data. The simulation studies are based on stylized and simplified data but allow for the models to be compared to a known true effect. They also allow for the sensitivity of the findings from different models to deviations from optimal assumptions to be examined and compared. The actual data studies provide for relative statements to be made and also enable the examination of year-to-year consistency in estimates. We discuss the findings in the context of the potential outcomes model, and the demands that it places on teacher effect estimates. This context greatly simplifies the discussion, as a number of the studies cited provide highly detailed accounts of assumptions concerning a number of parameters that are unobserved correlates of student learning that may or may not be relevant to teacher effect estimation. The potential outcomes framework provides some clarity regarding whether and how these may affect estimation of teacher effectiveness. As the studies are a wide mix of data and model types, we review them in chronological order. Some of the studies were conducted for the purpose of school accountability rather than teacher accountability, but the inferences can be reasonably extended to teacher effects, despite some differences in how confoundedness may emerge for teacher and school effects.

Tekwe et al. (2004) examined VAMs for school accountability (rather than teacher performance) and compared several models: a school fixed effects model with a change score regressed on school dummy variables; an unconditional hierarchical linear model (HLM) of a change score regressed on school random effects; a conditional hierarchical linear model of the change score regressed on student covariates for minority status and poverty, school covariates for mean pretest and percentage of poverty, and a school random effect; and a layered mixed effects model of the type developed for EVAAS. The authors found that the layered and fixed effects models were highly correlated; that the conditional HLM was different from the fixed effect model and the layered model, owing largely to the inclusion of student covariates; and that the unconditional HLM was very similar to the fixed effect model and layered model. This study makes important claims about the relationships among various fixed and random effects specifications, but the true model not being known, it cannot claim which one is better.

McCaffrey et al. (2004) compared the performance of multiple random effects or mixed effects models on simulated correlated covariates data that reproduced the complex nesting of students in multiple teachers with different cross-nestings of students over time. The models included a standard cross sectional hierarchical linear model (HLM) predicting achievement from covariates including prior scores; a serial cross-section design using gains as the dependent variable; a cross-classified linear growth HLM, where the cross-classification takes account of the nesting in multiple teachers over multiple years in the data; a layered (EVAAS/MRM) model; and a "general" model that incorporates covariate data and estimable teacher effect persistence parameters into the layered model framework. The other models were thus characterized as variously restricted versions of the general model. The authors simulated a small sample of 200 students in 10 classes of 20 for four grade levels and two schools using the general model and then comparing the estimates from each of the other models. Three different scenarios were tested based on ignorability (no violation; differential assignment of fast and slow learners in

classes; same differential assignment over schools), but teachers were randomly assigned to classes. The cross-classified and layered models performed better under all three scenarios, having higher correlations with the true estimates. In addition, the authors also demonstrated these models on actual data from five elementary schools in a large suburban school district, using the general model and the layered model and a single covariate (subsidized lunch eligibility), finding that their correlations ranged from .69 to .83. In both the simulated and actual findings, the comparisons were expressed solely as correlations, and it is not clear how many teachers would have been misclassified as ineffective under the various models and scenarios examined.

Guarino, Reckase, and Wooldridge (2012) conducted the most diverse simulation study to date, to test the sensitivity of a set of six VAMs to different student and teacher assignment scenarios. For the assignment of students to classrooms, two types of mechanisms were considered: dynamic (based on the previous year's test score) and static due to potential learning or actual learning at the time of enrollment. Teacher assignment to the classroom included a random scenario, and scenarios of systematic assignment with high-performing students assigned to high-performing teachers, or alternatively, low-performing teachers. The six models included a pooled ordinary least squares model with achievement in time *t* as the outcome (DOLS); the Arellano and Bond (1991) IV model, using a twice-lagged achievement assumed to be uncorrelated with current and lag errors as an instrument; a pooled OLS gain score, which is similar to the DOLS except that lagged achievement was forced to have a coefficient of 1; an "average residual" model that is similar to the teacher fixed effect model with covariates rather than fixed effects but excludes the teacher dummy variables (instead, the teacher estimates are obtained from a second-stage averaging of the student residuals to the teacher level); a student fixed effect model using the gain score as the outcome; and a variation on the gain score model with random effects for teachers. Using Spearman rank order correlations, Guarino, Reckase, and Wooldridge found that the DOLS was the best performing model. The DOLS was the only model that did not incorporate differencing on the lefthand side, and controlled directly for ignorability by incorporating previous achievement with a freely estimable parameter. A random effects variation on this model was not tested. Non-random groupings of students had minor effects on the results, but non-random teacher assignment had a deleterious effect on all of the estimates, particularly for heterogeneous student groupings with negative assignment to the teacher.

Finally, Schochet and Chiang (2010) used a simulation based on the decomposition of variance into student, classroom, teacher, and school components to compare OLS and random effects estimates of teacher effects using error rate formulas. These error rate formulas estimate the probability that a teacher in the middle part of the distribution of teachers will be found highly ineffective, and that a teacher who is truly ineffective will be considered no worse than average. Variance estimates for the decomposition were based on reported results in the literature. The authors demonstrated that under typical sample sizes, error rates were 10% for the OLS and 20% for the random effects models.

As of yet, no study has attempted to compare both fixed and random effects models of multiple types; Guarino, Reckase, and Wooldridge (2012) limited their examination to one random effects model, and it used a change score (imposing a coefficient of unity on the lag or pretest) rather than the achievement score itself. In this study, we compare a set of nine fixed and random

effects models, using rank order correlations and practical criteria such as misclassification, to answer the following questions:

1. Are the rankings of teachers from VAMs highly correlated with the ranking of "true" teacher effects under ideal (i.e., minimally confounded) conditions?

2. How accurately do the teacher effect estimates from VAMs categorize a teacher as ineffective, and what proportion of teachers would be misclassified?

3. How accurately do VAMs rank and categorize teachers when SUTVA is violated and classroom variance accounts for a proportion of the teacher effect?

4. How accurately do VAMs rank and categorize teachers when ignorability is violated and student effects are correlated with classroom, teacher, and school effects?

5. How similar are rankings of VAM estimates to each other?

6. How consistent are VAM estimates across years?

For each research question, we examined the relative performance of each VAM. To answer questions 1–5, we used simulation data, which gave us the advantage of having a known "true" effect. To answer questions 5 and 6, we used actual data collected from North Carolina schools.

Methods

We subjected nine models developed from the seven models described previously to examination and comparison using two types of simulation data and actual data collected in North Carolina. The nine models are summarized in Table 1. Three variations on the nested random effects model were estimated. The HLM2 was a two-level model in which β_s, β_{w-2}, and $u_s$were assumed to be 0; that is, only one pretest was used and there was no school effect estimated. The HLM3 was a three-level model in which β_{w-2} was assumed to be 0 (only one pretest). The HLM3+ was a three-level model in which all of these parameters were free. The five types of fixed effects models—student, teacher, student IVE, teacher IVE, and the DOLS—were all included exactly as described above. Finally, while an effort was made to examine the most widely known EVAAS model, the multivariate response model (MRM), it was ultimately not incorporated into this study due its high computational demands given the size of the data sets used. Instead, the URM, which is currently being implemented on a statewide basis in several states including North Carolina, was examined. We briefly revisit this limitation in the discussion. A fixed number of years (three) of matched student and teacher data was available for estimation in both the simulation and actual data, with the actual data also having two additional years of student end-of-grade performance matched to the students but not matched to the teachers from those years. We first discuss the data generation process for the simulations, and then the actual North Carolina data used. Finally, we discuss the methods used to compare the ten approaches.

Table 1. Summary of Value-Added Models

<table>
<tr><th>Model</th><th>Type</th><th>Cross-sectional or panel</th><th>Time invariant covariates</th><th>Lagged outcomes (pretests)</th><th>Teacher effect parameter</th><th>School random effect</th></tr>
<tr><td>HLM2</td><td rowspan="3">Nested random effects</td><td rowspan="3">Cross-sectional</td><td rowspan="3">Yes</td><td>1 in same subject</td><td rowspan="3">Teacher random effect (Empirical Bayes shrinkage estimator)</td><td>No</td></tr>
<tr><td>HLM3</td><td>1 in same subject</td><td rowspan="2">Yes</td></tr>
<tr><td>HLM3+</td><td>2 in each of two subjects</td></tr>
<tr><td>SFE</td><td rowspan="5">Fixed effects</td><td>Panel</td><td>Differenced to zero</td><td>None (all outcomes as Dependent Variable)</td><td>Mean of within-teacher residuals</td><td rowspan="5">N/A</td></tr>
<tr><td>TFE</td><td>Cross-sectional</td><td>Yes</td><td>1 in same subject</td><td>Teacher fixed effect (dummy variable)</td></tr>
<tr><td>SFEIV</td><td rowspan="3">Panel (two periods)</td><td rowspan="3">Differenced to zero</td><td rowspan="2">Once lagged as endogenous predictor; twice lagged as instrument (same subject)</td><td>Mean of within-teacher residuals</td></tr>
<tr><td>TFEIV</td><td rowspan="2">Teacher fixed effect (dummy variable)</td></tr>
<tr><td>DOLS</td><td>1 in same subject</td></tr>
<tr><td>URM</td><td>Hybrid Random Effects and Fixed Effects</td><td>Cross-sectional</td><td>No</td><td>2 in each of two subjects used to calculate composite</td><td>Teacher random effect (EB shrinkage estimator)</td><td>No</td></tr>
</table>

Data Generation Process

Analyses were conducted based on simulations of "typical" student, teacher, and school data, which enabled us to control the data generation process, thereby providing knowledge of "true" estimates of each simulated teacher's effectiveness against which to compare the results of each VAM. We assumed that the purpose of the VAM is to inform comparisons among teachers in a statewide evaluation system. We made several simplifications to make the data generation process and estimation more tractable. First, school and LEA membership of both teachers and students were fixed and consequently neither students or teachers could change school or LEA; second, we estimated models for the typical elementary school organization in which students and teachers are assigned to a classroom where instruction in several subjects occurs; third, we used only 5th grade, to assess teacher effectiveness; fourth, we created datasets with complete data (no missing data); fifth, only one subject was used as an outcome, though in some cases two subjects were used as control variables; sixth, the simulated data consisted of multiple districts but was much smaller than the population of districts in states, including North Carolina, which is the relevant reference since we employed actual data from there. However, enough teacher records were generated in order to ensure the teacher effect estimates were not subjected to problems commonly found in small samples (see below for sample sizes).

Two different simulations were used, each for answering a different question regarding violations of assumptions. In the first, the data generation process was developed via *variance decomposition* of each student's score into each level of schooling (student, classroom, teacher, and school). In this simulation, each teacher effect was homogenous across all students that teacher taught. This simulation focused on the effect that unaccounted-for classroom variance had on the ranking of the teacher effects from the nine models. In this simulation, a student covariate was included, but its correlation with the teacher effect was modest and had almost no effect on the estimates. Consequently, the data generation process of the second simulation created correlated student, classroom, teacher, and school covariates. This data generation process yielded a *heterogeneous teacher effect*, with teacher effects that varied across students having the same teacher, but having the desired level of correlation with the student, classroom, and school covariates. To infer teacher-level effectiveness, the mean level of the teacher effect was used as the "true" estimate. Both of the simulations were multilevel with random errors imputed to vary only within the appropriate level (e.g., classroom errors did not vary within classroom).

Variance Decomposition Simulation

We devised this simulation to answer questions 1, 2, and 3. In this simulation, each level of the data generation process—student, classroom, teacher, school, and LEA—was associated with a pre-specified variance component that represented that level's proportion of overall variance in the outcome. To ensure the data were as realistic as possible, the nesting structure for the simulation was an MMMC design. In this design, multiple cohorts of students were observed over a panel of three grades (3rd to 5th) over a period of three years for each cohort. The data generated with this process were uniquely designed to answer the question of the effect of violations of SUTVA (question 3), which may occur regardless of student-teacher assignment (ignorability) and thus could be examined without regard to correlations between student background and teacher effectiveness. Consequently, during these three grades, simulated

students were randomly sorted each year and assigned to teachers in these different randomly ordered groupings. The variance components for each level or type of input were then converted to standard deviations and multiplied by simulated standard normal random variables to identify each input's contribution to student learning. The statewide mean μ_{jw} for each of six standard normal $\sim N(0,1)$ test scores—over three grade levels and two subjects—was specified and then added to the subject-area-specific but time-invariant effects for student (ϕ_{ik}), classroom (ϕ_{ck}), teacher (ϕ_{jk}), school (ϕ_{sk}), and LEA (ϕ_{dk}) effects created via the variance decomposition, as well as a random or measurement error component (r_{icjsdkw}), to arrive at the total score for each student in each grade level and subject, as follows (with i = student, c = classroom, j = teacher, s = school, and k = subject, all defined as above, adding d for district):

$$(8) \qquad Y_{\text{icjsdkw}} = \mu_{jw} + \phi_{ik} + \phi_{ck} + \phi_{jk} + \phi_{sk} + \phi_{dk} + r_{\text{icjsdkw}}$$

The true teacher effect was the subject-area specific teacher input to student learning entered into model 8 (ϕ_{jk}). Each teacher was assumed to teach one group of 17–23 students (randomly determined) in any given year and to teach a common group of students in two subjects (math and reading). Therefore, for any cohort and subject area, the peer effect and teacher effect could not be distinguished in VAM estimation, and thus multiple cohorts were needed to separately estimate a classroom and teacher effect at this schooling level. We generated two cohorts of students. In each simulated cohort, there were 99,252 records generated, consisting of 16,542 students taking three end-of-grade exams in each of two subjects. A total of 833 teachers were simulated across 184 schools in 14 districts. The amount of variance attributed to the classroom was varied, taking on a value of 0 or 4%.

Heterogeneous Fixed Effects Simulation

The second simulation, designed to answer questions 1, 2, and 4, made use of a correlation matrix decomposition procedure that allowed us to feed into the simulation the desired level of correlation between two student covariates, one classroom and teacher covariate in each of three grade levels, and a school covariate (Vale & Maurelli, 1983). The resulting "effects" at each level were heterogeneous because they varied within their respective units; e.g., the classroom effect varied within the classroom. However, the procedure provided a high degree of control over the level of correlation between covariates for each level of schooling, necessary in order to generate non-randomness in the assignment of students to teachers, and to therefore be able to answer questions related to ignorability (question 4). The correlated covariates included one time-invariant student background effect for each of two subjects (μ_{ik}); one classroom effect for each grade level (of three) for a specific subject (μ_{ckw}, with w = 1, 2, 3); one teacher effect for each grade level for a specific subject (μ_{ikw}); and one grade- and subject-invariant school effect (μ_s; the school effect subsumed the district level effect). The correlations between classroom effects or between classroom and teacher effects were set at .20; between teacher effects across the three grades at .50; and between classroom or teacher and school effects at .20. Correlation between student effects and all others was varied, being either -0.20 or .20.

Similar to the variance decomposition simulation, a set of random effects representing residual variance at each level was then simulated at their respective levels (e.g., classroom residuals did not vary within classroom) and multiplied by standard deviations derived from pre-specified

variances. These included η_{ikw} for student, η_{ck} for classroom, η_{kw} for teacher, and η_{sk} for school. A subject- and grade-specific state grand mean (μ_{kw}) and residual ($e_{icgsdkw}$) were also estimated. All of these fixed and random effects were added together to produce the total achievement score for each student, as follows:

$$(9) \qquad \text{Yicjskw} = \mu_{kw} + \mu_{ikw} + \mu_{ckw} + \mu_{jkw} + \mu_s + \eta_{ikw} + \eta_{ck} + \eta_{jk} + \eta_{sk} + e_{icgsdkw}$$

The teacher effect, for the purposes of identifying a true value and estimating teacher effects from each VAM, was the teacher-level mean of the heterogeneous fixed teacher effect μ_{jkw}. This design had a much more parsimonious structure than the variance decomposition design, with peer groups advancing to the next grade level together rather than being re-sorted within cohort from year to year. We simulated 40,000 students in 2,000 classrooms, with 2 classrooms per teacher (representing two cohorts of students) and 1,000 teachers.

Calibration of Inputs

The inputs to the simulation that needed to be representative were the proportions of variance at each level, and we used two sources of information to justify them. First, we examined actual NC data for the grade levels in question. In elementary school, the math decomposition showed that little more than 10% of the variance was between teachers, with about 80% between students and the remainder between schools; reading was similar with 9% variance between teachers and 81% between students. We confirmed these inputs, to the extent possible, using the Nye, Konstantopolous, and Hedges study of variance decomposition (2004; refer to the authors' findings as well as Table 1), which suggested that teacher variance around 11% is consistent with the norm, though the grade levels examined (1st through 3rd) were lower than the grade level used in the current study (5th).

Number of Simulations

We ran 100 simulations of the designs specified above. This number of simulations is low relative to what is recommended in general for simulations, which is normally in the thousands. However, some concessions were required in order to keep the project manageable (each iteration of the 100 simulations required several hours to an entire day to process). There were some design components that helped to facilitate the use of a smaller number of simulations. First, we were not conducting hypothesis tests, but comparing estimands in each model with the "true" effect using a number of criteria (discussed in the next section). Second, a larger number of simulations is generally used in order to smooth out the variability between simulations imposed by measurement or random error. Alternatively, this could be done simply by minimizing the proportion of variance attributed to measurement error. Therefore, in these simulations, the amount of measurement or random error was specifically constrained to a fixed proportion of the variance (one percent1%). A sensitivity test was conducted to determine if 100 simulations was sufficient, comparing the results to a versions with 1,000 simulations. When measurement or random error was sufficiently low, the findings for 100 simulations were nearly identical to the findings for 1,000 simulations.

Actual NC Data Analysis

The second analysis was conducted on actual North Carolina data collected between 2007–08 and 2009–10, with some test score data also available from 2005–06 and 2006–07. This analysis was used to answer questions 5 and 6. Both math and reading end-of-grade standardized exam scores were used. While no "true" effect is known in this analysis, the data are the true North Carolina student performance data and not simplified as in the simulated data. In addition to the lagged scores or pretests as specified in the model, all relevant and commonly available student, peer, and school characteristics were incorporated into the analysis. These included student race/ethnicity, subsidized lunch, limited English proficiency, disability and academic giftedness, within and between-year movement, under-age and over- age indicators, previous years' peers' average standardized exam score, and an indicator of the year. Race/ethnicity and year were not included in the student fixed effects model or the two IVE models. In selected models (excluding the TFE and TFEIV), classroom covariates (class rank relative to 75th percentile in limited English proficiency, disability or giftedness, free lunch eligibility, and overage) and school covariates (proportion by race/ethnicity, total per-pupil expenditures, percent subsidized lunch eligible, violent acts per 1,000 and suspension rates in previous year, and enrollment) were included. No covariates at any level were entered into the URM. No teacher characteristics were included in the data because these teacher characteristics could explain the teacher effect that we actually wanted to estimate. The data used in this study consisted of all student records in 5th grade in North Carolina public schools matched via rosters to their teachers during three focal years. If the student has multiple teachers, records were weighted according to the proportion of the school year shared with each teacher.

For the consistency analysis, which required two sequential within-year estimates for each grade level, there were limitations to the amount of information available for the models that required multi-year panels to estimate (the SFE, SFEIV, TFEIV, and DOLS). For the SFE, SFEIV, and TFEIV, two sequences of three years' data were required for each model. However, among all time-varying covariates, only test score data were available prior to 2007–08 (giving us only three years of complete data), and therefore no time-varying covariates could be included in these models (differencing eliminates the time invariant covariates). For the DOLS, the panel was estimated with two years only, allowing for the inclusion of time-varying covariates. There were 503,370 student records in 5th grade math (8,826 teachers) and 728,008 student records in 5th grade reading (9,402 teachers).

Comparison Criteria

Three criteria were used to compare the absolute performance of each VAM on estimating the true teacher effects in the simulated data. Two of these and a different third criterion were used to assess relative performance of the models using actual NC data. First, Spearman rank order correlation coefficients, a non-parametric measure capturing the association between the rankings of two variables, was estimated for each pairing of a VAM with the true effect (simulation only) and with each other VAM (simulation and actual). For the simulated data, the estimates in each simulation needed to be combined into a single point estimate, which required a Fisher z transformation; the mean of this z-transformed correlation was calculated, and then back-transformed using the hyperbolic tangent function. High-performing VAMs have relatively higher Spearman coefficients.

Second, we calculated the percent agreement on the lowest 5% of the teacher quality distribution. The teachers in the bottom 5% of the distribution under each version of the teacher effect (the "true" effect in the simulation or from each VAM in both the simulated and actual) were identified. In the simulated data analysis, teachers' true and estimated scores agreed if they both ranked the teacher above the fifth percentile or they both ranked the teacher below the fifth percentile. The statistic was the proportion of all teachers with agreement. In the actual data, teachers' scores on any two methods agreed if the scores were both observed and were both above the fifth percentile or below. High-performing VAMs have relatively higher levels of agreement. Due to the normal distributions used in the data generation processes for the simulations, the findings for teachers in the 95th percentile were nearly identical. We chose this approach to correspond with a likely policy use of VAMs: to identify the lowest performing teachers.

Third, we examined the false identification of ineffective teachers in the simulated data only. For this analysis, special focus was placed on identifying a teacher who is actually relatively effective as ineffective based on their VAM score, due to the significant consequences that teachers and states may face under high stakes evaluation systems. We assumed a cutoff of -1.64 standard deviations from the mean teacher score, which is consistent with a finding of 5% of teachers being ineffective. First, we identified those teachers above the cutoff for ineffectiveness on the "true" measure. Then we identified those teachers who were below the cutoff on the estimated teacher effect. The teachers who satisfied both conditions were considered false positives or falsely identified as ineffective. This approach is a combination of the false positive/false negative methods used by Schochet and Chiang (2010). High-performing VAMs have relatively low proportions of false positives. Due to the normal distributions of the simulated data, we can assume that findings about falsely identifying a teacher as highly effective when he or she is not would be very similar. We also calculated the mean true score for teachers falsely identified as ineffective, and the number of teachers in North Carolina who would be affected by these findings. Actual data estimates for this comparison were not possible.

Fourth, we examined the year-to-year reliability in the VAMs in the actual NC data. For this criterion, the teacher estimates were obtained for each of two years individually. For the SFE, SFEIV, and TFEIV models, this required a substantial simplification of the models due to limitations in the actual NC data; further, for the DOLS no reliability analysis was possible, given these same limitations. Each teacher effect distribution on the eight remaining VAMs was divided into quintiles in each of the two years, and then each of these quintile classifications was cross-tabulated. If reliability were high, and allowing for some year-to-year variability including improvement, the teachers would have tended to fall along the diagonal where the quintiles were equal or roughly equal, with some off-diagonals suggesting an allowable amount of error and with the above-diagonal proportions slightly greater, allowing for improvement. If teachers did not fall along the diagonal, we could not tell which part would be due to estimate reliability and which part would be due to actual teacher improvement or change. We focused on three characteristics of the cross-tabulations: the proportion of teachers on the diagonal—that is, those teachers who were in the same quintile in each year—and the proportions of teachers in the most extreme "switchers" groups—those who were in the lowest quintile one year and the highest the next; or in the higher one year and the lowest the next. This method or one similar to it has been used by Sass (2008) and Goldhaber and Hansen (2008).

Results

We compared nine models' performance on a set of criteria that together were used to answer the six questions regarding rank ordering and identification of ineffective teachers with and without violations of potential outcomes assumptions, consistency across VAMs, and year-to-year reliability of VAMs. In reporting the results, we focus on the criteria and summarize the results into answers to the questions in the discussion section that follows.

Spearman Rank Order Correlations

Assessing performance by rank order correlations with the "true" effect assuming no classroom level variance (0% classroom variance), the best-performing VAM was the HLM3+ with three VAMs closely following in order: URM, SFE, and HLM3 (Table 2). The increase in the classroom proportion of variance for testing the influence of SUTVA and confoundedness reduced the Spearman rank order correlations of all models with the true effect (Table 2). The violation of SUTVA implied by 4% of variance at the classroom level did not affect the relative ranking of the VAMs on this criterion. The HLM3+ was highest at .955 at 0% classroom variance and remained highest at 4% classroom variance (.864). The HLM2, TFE, and DOLS were nearly equal (.909 and .822, respectively), as were the SFEIV and TFEIV (.893 and .808, respectively). The classroom variance simulated in this analysis at 4% should be considered reasonable, given the analysis of Schochet and Chiang (2010).

Table 2. Spearman Rank Order with True Effect, Simulated Data

Model	By % variance at classroom level		By correlation between student, classroom, teacher, and school fixed effects*	
	0% Classroom variance	4% Classroom variance	ρ = -.20	ρ = .20
HLM2	0.909	0.822	0.716	0.662
HLM3	0.934	0.844	0.796	0.746
HLM3+	0.955	0.864	0.771	0.755
SFE	0.941	0.851	0.648	0.628
TFE	0.909	0.822	0.077	0.059
SFEIV	0.893	0.808	0.562	0.526
TFEIV	0.893	0.808	0.062	0.048
DOLS	0.909	0.822	0.003	-0.007
URM	0.946	0.856	0.660	0.670

*Correlation (ρ) shown is between student and classroom, teacher and school; correlation between classrooms or classroom and teacher is .20; correlation between teachers is .50; correlation between classroom or teacher and school is .20.

When the strong ignorability of assignment (confounded assignment) was violated, there was substantial variation (Table 2) in the Spearman rank order for either moderate positive or negative correlation between the student covariate and the classroom, teacher, and school covariates, with two random effects models, the HLM3 (.796 and .746, respectively) and HLM3+(.771 and .755, respectively) being the top performers, followed by the HLM2 (.716 and .662, respectively), URM (.660 and .670, respectively), SFE (.648 and .628, respectively), and SFEIV (.562 and .526, respectively), but with the TFE, TFEIV, and DOLS very low. In the rank order correlation with optimal conditions, SUTVA violations, and confounded assignment, the HLM3 and HLM3+ were consistently the highest performing VAMs, and several models, including the four fixed effects and DOLS VAMs, performed much worse than the others.

Table 3. Spearman Rank Order Matrices of Value-Added Models, Actual Data

5th grade math

Model	HLM2	HLM3	HLM3+	SFE	TFE	SFEIV	TFEIV	DOLS	URM
HLM2	1.000	0.872	0.845	0.757	0.944	0.850	0.686	0.866	0.916
HLM3	0.872	1.000	0.970	0.657	0.822	0.741	0.594	0.750	0.799
HLM3+	0.845	0.970	1.000	0.708	0.796	0.726	0.661	0.729	0.828
SFE	0.757	0.657	0.708	1.000	0.727	0.812	0.759	0.680	0.831
TFE	0.944	0.822	0.796	0.727	1.000	0.808	0.709	0.904	0.891
SFEIV	0.850	0.741	0.726	0.812	0.808	1.000	0.642	0.729	0.834
TFEIV	0.686	0.594	0.661	0.759	0.709	0.642	1.000	0.741	0.830
DOLS	0.866	0.750	0.729	0.680	0.904	0.729	0.741	1.000	0.868
URM	0.916	0.799	0.828	0.831	0.891	0.834	0.830	0.868	1.000

5th grade reading

Model	HLM2	HLM3	HLM3+	SFE	TFE	SFEIV	TFEIV	DOLS	URM
HLM2	1.000	0.927	0.876	0.642	0.813	0.758	0.549	0.733	0.853
HLM3	0.927	1.000	0.948	0.591	0.755	0.706	0.491	0.677	0.789
HLM3+	0.876	0.948	1.000	0.664	0.713	0.680	0.574	0.645	0.836
SFE	0.642	0.591	0.664	1.000	0.530	0.731	0.595	0.488	0.705
TFE	0.813	0.755	0.713	0.530	1.000	0.636	0.606	0.861	0.765
SFEIV	0.758	0.706	0.680	0.731	0.636	1.000	0.511	0.564	0.720
TFEIV	0.549	0.491	0.574	0.595	0.606	0.511	1.000	0.674	0.750
DOLS	0.733	0.677	0.645	0.488	0.861	0.564	0.674	1.000	0.771
URM	0.853	0.789	0.836	0.705	0.765	0.720	0.750	0.771	1.000

On the actual NC data (see Table 3 containing two correlation matrices), the rank order correlations between the VAM estimates varied considerably in both math and reading, from .970 to .642 for mathematics and .948 to .488 for reading. In both subjects, the URM was most highly correlated with the other models, averaging .850 and .774, respectively. The TFEIV was the least highly correlated with the other models, averaging .793 and .594, respectively. The two most highly correlated VAMs were HLM3 and HLM3+, with .970 for mathematics and .948 for reading. There was a tendency for the random effects models to be highly correlated with each

other and the URM and TFE models. The TFE model was highly correlated with the HLM2 (.944 for 5th grade math and .813 for 5th grade reading) and DOLS (.904 for 5th grade math and .861 for 5th grade reading) but not with the other fixed effects models. The fixed effects models did not exhibit an overall tendency to be highly correlated with each other or to be more highly correlated with each other than with the random effects models. Overall, it appears that the choice of a VAM model over some others can yield quite different rank orderings of the teacher effect estimates. It is important to note that higher correlations between the VAM model estimates from the actual data do not imply that they recover the "true" teacher effect estimates more consistently because the models may be reproducing a similar bias.

Agreement on Classification in Fifth Percentiles

The agreement on classification provides an indication of the extent to which the VAMs agree with the true effect or each other in terms of identifying the lowest performing 5% of teachers in the state. This criterion is quite important when the teacher effect estimates are to be used for teacher evaluations with consequences, since there are significant costs associated with falsely identifying an average teacher in the lowest performing group or falsely identifying a low-performing teacher in the "acceptable" range of performance. Nearly all of the VAMs performed very well in the absence of assumption violations, with between 97.7% and 96.3% agreement on the bottom 5% and top 95%, which is less than a 1.5% difference (Table 4). In the test of the SUTVA violation with 4% of the variance at the classroom level, the VAM exhibited lower agreement rates, about 95%–96%, with the difference between the models much less, having a range of only 0.82. The HLM3+ was the highest, with 97.71% agreement with zero classroom variance, and it remained the highest with 4% classroom variance (96.01%). Nevertheless, all of the coefficients were very similar.

Table 4. Percent Agreement with True Effect, Simulated Data

Model	By % variance at classroom level		By correlation between student, classroom, teacher, and school fixed effects*	
	0% Classroom	4% Classroom	ρ = -.20	ρ = .20
HLM2	96.65	95.37	94.25	93.79
HLM3	97.21	95.74	95.04	94.58
HLM3+	97.71	96.01	94.78	94.72
SFE	97.33	95.75	93.56	93.52
TFE	96.65	95.36	90.93	90.75
SFEIV	96.31	95.20	92.98	92.78
TFEIV	96.33	95.19	90.74	90.69
DOLS	96.62	95.35	90.48	90.50
URM	97.44	95.81	93.70	93.92

*Correlation (ρ) shown is between student and classroom, teacher and school; correlation between classrooms or classroom and teacher is .20; correlation between teachers is .50; correlation between classroom or teacher and school is .20.

In the test of the confounded assignment, the level of agreement was reasonably high with all models at or above 90% agreement in the positive assignment and negative (compensatory) assignment scenarios. The HLM3 and HLM3+ were the highest agreement models (for the positive assignment, 95.04 and 94.78, respectively), followed by the HLM2, URM, SFE, and SFEIV (for the positive assignment, 94.25, 93.70, 93.56, and 92.98, respectively). Three consistently lower performers were the TFE, TFEVI, and DOLS in the positive assignment (90.93, 90.74, and 90.48, respectively), with the negative assignment following the same pattern. There was a more sizeable gap between the higher and lower ranking models than for the variance decomposition findings, and the direction of the correlation did not alter the pattern.

With the actual NC data, the agreement between the VAMs was quite high with all models, averaging from 94%–95% agreement with each other for mathematics and reading (Table 5). There was a tendency for the random effects VAMs to be in greater agreement with each other, and for the fixed effects VAMs (including the DOLS) to be in greater agreement with each other, with lower agreement across type. This tendency was not as great in math—the percentage of agreement in each partition of the matrix was very similar—but was obvious in reading.

Table 5. Percent Agreement Across Models, Actual Data

5th grade math

Model	HLM2	HLM3	HLM3+	SFE	TFE	SFEIV	TFEIV	DOLS	URM
HLM2	100.00	95.90	95.49	94.67	96.12	95.35	93.25	95.08	96.69
HLM3	95.90	100.00	98.03	93.93	94.36	94.37	92.88	93.88	95.04
HLM3+	95.49	98.03	100.00	94.49	94.08	93.91	93.35	93.65	95.47
SFE	94.67	93.93	94.49	100.00	95.07	95.34	95.20	95.05	95.89
TFE	96.12	94.36	94.08	95.07	100.00	96.08	94.94	97.45	95.40
SFEIV	95.35	94.37	93.91	95.34	96.08	100.00	93.80	95.47	95.03
TFEIV	93.25	92.88	93.35	95.20	94.94	93.80	100.00	95.09	95.15
DOLS	95.08	93.88	93.65	95.05	97.45	95.47	95.09	100.00	95.26
URM	96.69	95.04	95.47	95.89	95.40	95.03	95.15	95.26	100.00

5th grade reading

Model	HLM2	HLM3	HLM3+	SFE	TFE	SFEIV	TFEIV	DOLS	URM
HLM2	100.00	97.28	96.41	93.86	94.11	94.14	92.57	94.07	96.04
HLM3	97.28	100.00	97.86	93.71	93.73	93.73	92.49	93.78	95.26
HLM3+	96.41	97.86	100.00	94.37	93.50	93.54	93.15	93.57	95.83
SFE	93.86	93.71	94.37	100.00	94.75	95.62	95.12	94.17	94.47
TFE	94.11	93.73	93.50	94.75	100.00	95.06	94.73	96.90	93.90
SFEIV	94.14	93.73	93.54	95.62	95.06	100.00	93.95	94.13	93.89
TFEIV	92.57	92.49	93.15	95.12	94.73	93.95	100.00	94.68	94.09
DOLS	94.07	93.78	93.57	94.17	96.90	94.13	94.68	100.00	94.35
URM	96.04	95.26	95.83	94.47	93.90	93.89	94.09	94.35	100.00

False Positives: Average Teacher Identified as Ineffective

The third type of analysis assessed the extent of false positives; that is, how many teachers in the top 95% of the distribution would be falsely identified as bottom 5% performers. This criterion is relevant because several have proposed to use VAM estimates of teacher effectiveness to identify "ineffective" teachers as a step toward dismissal. False positives were examined on the simulated data only (Table 6, following page). In the variance decomposition simulation, at low levels of classroom variance (an absence of assumption violations), the HLM3+ (1.2% false positives), URM (1.3%), HLM3 (1.4%), and SFE (1.4%) performed the best; the other models were 1.7% or higher. To get a more concrete estimate of the breadth of the differences in model performance, assuming 9,000 5th grade teachers (the approximate number statewide in North Carolina), between 108 and 170 would be falsely identified as ineffective by the best and worst performing VAMs. In other words, the worst performing VAM would falsely identify 62 more 5th grade teachers as ineffective. The mean of the true z-score for these teachers was -1.43 for the HLM3+, the best performing VAM, and -1.30 for the worst performing VAMs, the SFEIV and TFEIV, which indicates that the teachers being falsely identified as ineffective by the worst performing VAMs were on average better performers; false identification of ineffectiveness casts a wider net in the worst performing models.

When the level of classroom variance was set at 4%, however, the relative performance advantage of all models declined somewhat, with all of the models demonstrating higher proportions of false positives (2.0%–2.4 % at 4% classroom variance). While these rates were seemingly modest, the number of teachers affected in each grade level and subject can be large, with up to 210 teachers misclassified under a scenario with 4% variance. The differences among the models, however, were modest, with a difference of 28 teachers at most.

With a heterogeneous fixed effect simulation, there was substantial variation between the models in the proportion of teachers misidentified as ineffective in the positive assignment scenario (Table 6, following page), with the HLM2, HLM3, and HLM3+ being the best performers (less than 3% misidentified), followed by the URM, SFE, and SFEIV misidentifying 3.1%, 3.2%, and 3.5%, respectively, and the TFE, TFEIV, and DOLS misidentifying more than 4%. The direction of correlation altered this pattern only slightly with only the HLM3 and HLM3+ incorrectly identifying less than 3% of the ineffective teachers followed closely by the HLM2, URM, and SFE. The number of teachers affected nearly doubled from the best to the worst performing models on this criterion, ranging from 221 (HLM3) to 436 (DOLS). Finally, for the worst performing VAMs, the TFE, TFEIV, and DOLS, the point estimates for the mean true effect of the misidentified teachers were actually above zero, meaning that the misclassified teachers included above-average teachers.

Table 6. False Positives: Teachers Falsely Identified as Ineffective, Simulated Data Only (Cutoff at -1.64 from Mean5% of Teachers Ineffective)

By % variance at classroom level

Model	0% Classroom variance			4% Classroom variance		
	%	No. of teachers affected*	Mean true Z score of affected teachers	%	No. of teachers affected*	Mean true Z score of affected teachers
HLM2	1.7%	155	-1.33	2.3%	210	-1.18
HLM3	1.4%	128	-1.37	2.2%	194	-1.22
HLM3+	1.2%	108	-1.43	2.1%	193	-1.24
SFE	1.4%	122	-1.40	2.0%	182	-1.26
TFE	1.7%	155	-1.33	2.2%	194	-1.23
SFEIV	1.9%	170	-1.30	2.3%	209	-1.18
TFEIV	1.9%	170	-1.30	2.4%	219	-1.15
DOLS	1.7%	153	-1.33	2.4%	219	-1.15
URM	1.3%	118	-1.41	2.3%	210	-1.18

By correlation between student, classroom, teacher, and school fixed effects^

Model	ρ = -.20			ρ = .20		
	%	No. of teachers affected*	Mean true Z score of affected teachers	%	No. of teachers affected*	Mean true Z score of affected teachers
HLM2	2.9%	259	-1.00	3.1%	277	-0.94
HLM3	2.5%	221	-1.14	2.7%	244	-1.06
HLM3+	2.7%	239	-1.09	2.7%	242	-1.08
SFE	3.2%	290	-0.90	3.2%	288	-0.88
TFE	4.6%	412	0.05	4.6%	415	0.06
SFEIV	3.5%	314	-0.78	3.6%	325	-0.73
TFEIV	4.6%	413	0.06	4.7%	420	0.05
DOLS	4.8%	436	0.12	4.8%	436	0.13
URM	3.1%	283	-0.93	3.0%	270	-0.95

*Assumes 9,000 5th grade teachers in North Carolina.
^Correlation (ρ) shown is between student and classroom, teacher and school; correlation between classrooms or classroom and teacher is .20; correlation between teachers is .50; correlation between classroom or teacher and school is .20.

Reliability

Using actual NC data, quintiles of teacher performance were estimated for two sequential years, then these results cross-tabulated at the teacher level. This cross-tabulation represented a pattern of mixing across the two years of teachers at each level of effectiveness in the first year. The cross-tabulations of the quintiles in 2007–08 and 2008–09 were summarized into two summary tables (Table 7, following page). The sum of the percentage of teachers on the diagonal of each cross-tabulation, with the teachers on the diagonal being those in the same quintile in each year and the sum of the percentage of teachers who either switched from the top quintile to the bottom

or from the bottom quintile to the top during the same interval were both included in Table 7 for both math and reading. For 5th grade teachers, the DOLS outperformed all others with 44.5% for math and 39.2% for reading, followed by the URM with 33.2% for math and 28.3% for reading. The lowest percentages of year-to-year quintile consistency were the HLM3 and HLM3+, virtually tying with 30% for math and 25% for reading.

Table 7. Consistency of Teacher Performance Rankings over Consecutive Years, Actual Data Only

Model	Same quintile both years		Switch from lowest to highest or highest to lowest	
	Math	**Reading**	**Math**	**Reading**
HLM2	32.5	25.7	1.8	4.6
HLM3	29.5	24.6	3.1	6.0
HLM3+	30.0	24.5	3.2	5.9
SFE	34.6	27.5	2.2	4.3
TFE	33.3	28.9	1.7	3.9
SFEIV	31.5	26.5	2.3	4.8
TFEIV	31.5	26.6	2.3	4.8
DOLS	44.5	39.2	0.2	0.8
URM	35.1	28.3	1.7	4.3

There was also some variation between the models in the percentage of teachers switching from one extreme quintile to the other extreme, and there was a clear difference between math and reading in the performance of these models, with about twice as many teachers switching in reading as in math in 5th grade. The DOLS was the best performer, with 0.2% switching in math and 0.8% in reading. The other models were more similar, but the HLM3 and HLM3+ had the highest percentages of switching in both reading and math.

Discussion

Using two simulations of student test score data as well as actual data from North Carolina public schools, we compared nine value-added models on the basis of four criteria related to teachers' effectiveness rankings: Spearman rank order; percentage of agreement on 5th percentile; false positives consisting of teachers who are not ineffective being misidentified as ineffective; and consistency of rankings within quintiles over two sequential years. Using these comparisons, we answer six questions that are pertinent to state policymakers and administrators who may be in positions to select a value-added model to obtain estimates of individual teachers' effectiveness generated from student test score data and to the teachers and principals who may be directly affected by them.

1. *Are the rankings of teachers from VAMs highly correlated with the ranking of "true" teacher effects under ideal (i.e., absence of assumption violations) conditions?*

   While all nine VAMs performed reasonably well on this test, four models were higher performers (the HLM3+, URM, SFE, and HLM3) than the other five.

2. *How accurately do the teacher effect estimates from VAMs categorize a teacher as ineffective, and what proportion of teachers would be misclassified?*

   While all models performed reasonably well on this test, four models were higher performers (HLM3+, URM, SFE, and HLM3) than the other five.

3. *How accurately do VAMs rank and categorize teachers when SUTVA is violated and classroom variance accounts for a proportion of the teacher effect?*

   For the accuracy of ranking when SUTVA is violated, the performance of all models was substantially reduced in comparison to the absence of assumption violations. In terms of relative performance, four models were higher performers (HLM3+, URM, SFE, and HLM3) than the other five. For the accuracy of categorizing teachers in the lowest 5% when SUTVA is violated, all VAMs performed equivalently.

4. *How accurately do VAMs rank and categorize teachers when ignorability is violated and student effects are correlated with classroom, teacher, and school effects?*

   For the accuracy of ranking when ignorability is violated, the performance of the VAMs was somewhat reduced in comparison to the absence of assumption violations. The relative performance of the VAMs varied substantially; two models were higher performers (the HLM3+ and HLM3) than the other seven in both the negative assignment and positive assignment scenarios. For the accuracy of categorizing teachers in the lowest 5% when ignorability is violated, all VAMs correctly classified more than 90% of the teachers with six models, the HLM3+, HLM3, HLM2, URM, SFE, and SFEIV, outperforming the other four.

5. *How similar are rankings of VAM estimates to each other?*

   For mathematics, the rankings produced by three VAMs, the URM, HLM2, and TFE, are more similar to all others (the average of all VAMs) than the other six models. For reading, the rankings produced by five models, the URM, HLM2, HLM3+, HLM3, and TFE, are more similar to all others than the other four VAMs.

6. *How consistent are VAM estimates across years?*

   The most consistent year-to-year VAM estimates in terms of placing the highest percentage of teachers in the same performance quintile are the DOLS and URM. In terms of consistency in producing the fewest highest to lowest or lowest to highest switchers, the DOLS is the best performing VAM, followed by the TFE, URM, and HLM2.

Clearly, the overall ranking of model performance depends on how the criteria are weighted. If performance of the models in the presence of violated ignorability is viewed as the most highly weighted criteria, three VAMs performed sufficiently poorly to appear to be risky choices for estimating individual teacher effectiveness—teacher fixed effects, teacher fixed effects with IV, and dynamic ordinary least squares. None of these three models performed well in either test for confounded assignments of students and teacher, and much research strongly suggests confounded assignment is frequently the case now. In the simulations violating SUTVA assumptions, these models seem to underperform relative to the others in the ranking but not in the identification of the 5% of poor performers. And neither did they underperform in the examinations of year-to-year consistencies. This conclusion may need to be tempered in the case of the DOLS because of the relatively high performance of that VAM in the simulations by Guarino, Reckase, and Wooldridge (2012), but their findings with respect to the teacher fixed effects VAM are consistent.

More research should be done examining the performance of the DOLS before a strong affirmative recommendation could be offered. Bearing in mind that the findings of the present study and Guarino, Reckase, and Wooldridge (2012) regarding the DOLS only overlap in examining rank correlations, we speculate that the DOLS may be a higher performer in the Guarino, Reckase, and Wooldridge study for a number of reasons. Differences in the data generation processes, combined with the choice of the authors not to examine a model with raw score as the outcome and a shrinkage estimator for the teacher effect may be the cause for this seeming disagreement. Teacher estimates shrunken by empirical Bayes were applied to the gain score, but the authors argued that with invariant class sizes in their design, the shrinkage estimator on the raw score would produce rankings equivalent to that for the DOLS. As a consequence, the DOLS estimates in the Guarino, Reckase, and Wooldridge study are equivalent to a random effects variant that they did not test. This is certainly consistent with the present study, as two of the simple nested random effects models (the HLM3 and HLM3+) were regularly among the highest performing models.

With the findings indicating that the TFE, TFEIV, and DOLS are risky, are there any that policymakers and administrators might wish to consider adequate? The answer to this question can only be answered by a definitive weighting scheme for the criteria, which should include an

assessment of the costs and consequences of the particular purposes for which the estimates will be used. The list of acceptable models could be quite different for estimates of teacher effectiveness that are used to identify teachers who may need additional professional development (low stakes) and those used to identify teachers for high stakes sanctions such as denial of tenure, dismissal, or substantial bonuses, with identification for additional observations with feedback and coaching and other positive benefits falling somewhere between. We believe the evidence suggests that four VAMs performed sufficiently well across the board to deserve consideration for low stakes purposes: the three-level hierarchical linear model with one year of pretest scores, the three-level hierarchical linear model with two years of pretest scores, the EVAAS univariate response model, and the student fixed effects model. The performance of each of these models was quite good in recovering the true effects and was quite similar. The performance of all was degraded by a violation of SUTVA—more so for the ranking and less so for agreement on classification in the bottom 5% and the false identification of ineffective teachers—but not as much by confounding. Also, quite relevant is the identification of the lowest fifth percentile, which could be used in low to medium stakes situations. In our opinion, these are relevant criteria for assessing adequacy of VAM for low stakes purposes.

We believe that the false positive analysis is particularly important when considering the adequacy of models for high stakes use. If SUTVA is substantially violated, about 350 5th grade teachers in a state the size of North Carolina could be identified for possible removal when their actual performance was not in the lowest 5% of teachers. The mean performance of these teachers is more than 0.6 of a standard deviation below the mean when the four higher performing models are used. When confounding occurs, the higher performing models would falsely identify approximately 220–290 5th grade teachers in a state about the size of North Carolina as in the lowest performing 5% of the distribution. For the four higher performing VAMs, these falsely identified teachers' average performance is at least 0.9 standard deviations below the mean. For many, this would seem to suggest that the teacher effectiveness estimates should at most be considered a first step in identifying ineffective teachers, rather than the method for identification of teachers for high stakes personnel actions. Using any VAM, even the highest performing ones, to identify teachers for high stakes consequences seems risky in our opinion.

It seems important to consider consistency as well when considering if any of the VAMs should be used for estimating individual teacher effectiveness. As earlier research points out, inconsistency in the estimates from year to year can undermine the credibility of the estimates, especially to those whose performance is being estimated (Amrein-Beardsley, 2008). The best performer in this regard, the DOLS, was a very low performer in the simulations; the EVAAS URM and student fixed effects performed somewhat better than the other two better performers, the HLM3 and HLM3+. However, all of the assessments are relative. It would be difficult to know whether the differences in VAM performance that we observed using the North Carolina data—3.2% switching from highest to lowest or vice versa rather than 1.7%—would affect credibility. The fact that these extreme switchers exist at all may be sufficient evidence to convince some policymakers and some teachers that no sufficiently consistent VAM exists. Further research should be conducted to better understand the correlates of the extreme quintile switching, in particular investigating the number of novice teachers that switch or the number of extreme switchers that have changed assignments, such as moving from one school to another or one grade to another.

Limitations and Implications

Limitations

This study had several limitations. First, a significant portion of the analysis was based on simulated stylized data. This was intended to address the absence of "true" measures of teacher effects in actual data. While these simplifications may suggest that real conditions would probably degrade the absolute performance of each model, we have not argued that this degrading of performing would be equivalent across all models, and therefore it is possible that more realistic conditions might influence the comparisons that we have made. For example, we did not simulate missing values, a problem typical of actual data that by design some of the models (e.g., the URM) may handle better than the others. Second, there was some necessary subjectivity in the choice and specification of models, including in the types of fixed effects models used and the covariates used in some models. Third, we were unable to estimate extensive simulations or actual data models for the EVAAS MRM, a controversial (Amrein-Beardsley, 2008) but widely published (Ballou, Sanders, & Wright, 2004; McCaffrey et al., 2004) model. While McCaffrey et al. (2004) suggested that this model performed similarly to a fixed effects model using small samples, our experience with a smaller variance decomposition sample than the one used in that study (144 teachers, rather than 833) suggests that the MRM performed poorly. A single simulation of 833 teachers with zero classroom variance, however, indicates that the MRM had very similar performance to the URM. Nevertheless, we cannot recommend the MRM, as its computational demands place it out of the reach of many state education agencies and scholars to estimate. Finally, the limited actual data, ranging over only three years of data in which students were matched to their teachers, made some of the analyses difficult to undertake and required some modifications to the models when multiple estimates were required for examining year-to-year consistency.

Despite these limitations, there are multiple strengths of this study. It is the first of its kind to use simulated variance decomposition and correlated fixed effect data specifically designed for testing both SUTVA violations and ignorability, respectively, as well as actual data. It is also the first of its kind to examine multiple random effects and fixed effects models, and it examined nine models, nearly twice that of any other study.

Implications

Value-added models for teacher effectiveness are a key component of reform efforts aimed at improving teaching and have been examined by this study and others. However, an interdisciplinary consensus on the methods used to obtain value-added teacher estimates does not exist, and many different models spanning multiple disciplines including economics and sociology have been proposed, as noted above. Further, several different approaches have been used to examine and compare models, and as this study demonstrated with just a handful of approaches, the "best" VAM may be dependent on the comparison approach. Nevertheless, when multiple approaches were used, trends did emerge that pointed to a few models that were on average better performers, and a handful that were almost universally poor. We suggest that one implication of this study is that multiple approaches are needed to get a fuller picture of the relative merits of each model.

This study showed that the most common VAMs produce teacher rankings having markedly high correlations with each other, particularly in models with no violations or only modest violations of SUTVA. With few exceptions, the Spearman rank correlation did not highly discriminate between models. Further, many of the observed rank correlations that we deemed inadequate in comparison to those of the best performing models would have been viewed as high in any other research setting. Much the same could be said for the percentage of agreement in being categorized in the bottom 5%, which had both very high levels of agreement and little in the way of discriminating power. On the other hand, when categorization into the bottom 5% was framed as a question of the rate and number of teachers falsely identified as ineffective—*false positives*—severe differences between some models emerged, particularly under the scenario of non-ignorable assignment. Further, the findings as presented actually understate the risk to education agencies from choosing the wrong model. For a negative assignment scenario, under the best model (the HLM3), 221 5th grade teachers would be misclassified, while under the worst model (the DOLS), 436 5th grade would be misclassified, a difference of 215 teachers. Recall, however, that it is equally true that the same number of teachers would be misclassified as highly effective, and that in both cases teachers who are misclassified as ineffective (or effective) are counterbalanced by an equally large group of teachers who should have been identified as ineffective (effective) but were not. This implies that the actual difference is four times the amount, for 860 teachers per grade level. And it further underscores the risks associated with using any of these models, as even the best model, which had a misclassification of 221 5th grade teachers as ineffective, produced an overall misclassification of almost 900 teachers per grade.

This points to the second implication that this study raises: the risks to these misclassified teachers and their students will depend upon the uses of these evaluations, particularly the stakes assigned. Based on our findings, we believe that four of the tested VAMs, the HLM3+, HLM3, URM, and SFE, can provide objective information about the effectiveness of individual teachers in affecting the test score changes of their students for low stakes purposes. The evidence in this study suggests that the use of any VAMs for high stakes purposes is quite risky, even for the best performing models. The evidence also suggests that several of the VAM models are likely to be less accurate in estimating the actual effects of teachers, including the TFE, TFEIV, and DOLS. Until additional research shows otherwise, these models should be considered risky even for low stakes purposes, although some of them performed well on some criteria even in this study.

References

Amrein-Beardsley, A. (2008). Methodological concerns about the Education Value-Added Assessment System. *Educational Researcher, 37*(2), 65–75.

Arellano, M., & Bond, S. (1991). Some tests of specification of panel data: Monte Carlo evidence and an application to employment equations. *The Review of Economic Studies, 58*, 277–298.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37–65.

Browne, W. J., Goldstein, H., & Rasbash, J. (2001). Multiple membership multiple classification (MMMC) models. *Statistical Modeling 2001 (1)*, 103–124.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *The Journal of Human Resources, 41*(4), 778–820.

Goldhaber D., & Hansen, M. (2008). *Assessing the potential of using value-added estimates of teacher job performance for making tenure decisions*. Washington, DC: National Center for Analysis of Longitudinal Data in Education Research.

Gordon, R., Kane, T. J., and Staiger, D. O. (2006). *Identifying effective teachers using performance on the job* (Hamilton Project Discussion Paper). Washington, DC: Brookings Institution.

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2012). *Can value-added measures of teacher performance be trusted?* (Working Paper #18). East Lansing, MI: The Education Policy Center at Michigan State University.

Hanushek, E. A. (1986). The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature, 24*, 1141–1177.

Harris, D. N. (2009). Teacher value-added: Don't end the search before it starts. *Journal of Policy Analysis and Management, 28*(4), 693–699.

Henry, G. T., Kershaw, D. C., Zulli, R. A., & Smith, A. A. (in press). Incorporating teacher effectiveness into teacher preparation program evaluation. *Journal of Teacher Education.*

Hill, H. C. (2009). Evaluating value-added models: A validity argument approach. *Journal of Policy Analysis and Management, 28*(4), 700–712.

Holland, Paul W. (1986) Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945–960.

Koedel, C., &Betts, J. R. (2011). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Education Finance and Policy, 6*(1), 18–42.

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.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research.* Cambridge, UK: Cambridge University Press.

Nye, B., Konstantopolous, S., & Hedges, L. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*(3), 237–257.

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

Raudenbush, S. W. & Willms, J. D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics, 20*(4), 307–335.

Reardon, S. F., & Raudenbush, S. R. (2009). Assumptions of value-added models for estimating school effects. *Education Finance and Policy, 4*(4), 492–519.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review, 94*(2), 247–252.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay and student achievement. *The Quarterly Journal of Economics, 25*(1), 175–214.

Rowan, B., Correnti, R., & 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*(8), 1525–1567.

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.

Sass, T. (2008). *The stability of value-added measures of teacher quality and implications for teacher compensation policy*. Washington, DC: National Center for Analysis of Longitudinal Data in Education Research.

Schochet, P. Z., & Chiang, H. S. (2010). *Error rates in measuring teacher and school performance based on student test score gains*. Washington, DC: Institute for Education Sciences.

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

Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal, 113*, f3–f33.

Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika, 48*, 465–471.

Wright, S. P., White , J. T., Sanders, W. L., & Rivers, J. C. (2010). *SAS EVAAS statistical models*. Cary, NC: The SAS Institute.

Contact Information:
Please direct all inquiries to Roderick A. Rose
rarose@email.unc.edu
Department of Public Policy and School of Social Work
The University of North Carolina at Chapel Hill
324E Tate-Turner-Kuralt
325 Pittsboro St., CB 3550
919.260.0542 (voice)

SERVECenter The University of North Carolina at Greensboro People · Possibilities · Promise
Carolina Institute for Public Policy
The University of North Carolina at Chapel Hill
The William & Ida Friday Institute for Educational Innovation
NC State University College of Education