Journal of Educational and Behavioral Statistics
http://jebs.aera.net

Models for Value-Added Modeling of Teacher Effects

Daniel F. McCaffrey, J. R. Lockwood, Daniel Koretz, Thomas A. Louis and Laura Hamilton

JOURNAL OF EDUCATIONAL AND BEHAVIORAL STATISTICS 2004 29: 67
DOI: 10.3102/10769986029001067

The online version of this article can be found at:
http://jeb.sagepub.com/content/29/1/67

Published on behalf of

American Educational Research Association

and

http://www.sagepublications.com

Journal of Educational and Behavioral Statistics
Spring 2004, Vol. 29, No. 1, pp. 67–101

Models for Value-Added Modeling of Teacher Effects

Daniel F. McCaffrey
J. R. Lockwood
RAND

Daniel Koretz
Harvard Graduate School of Education

Thomas A. Louis
Johns Hopkins Bloomberg School of Public Health

Laura Hamilton
RAND

The use of complex value-added models that attempt to isolate the contributions of teachers or schools to student development is increasing. Several variations on these models are being applied in the research literature, and policy makers have expressed interest in using these models for evaluating teachers and schools. In this article, we present a general multivariate, longitudinal mixed-model that incorporates the complex grouping structures inherent to longitudinal student data linked to teachers. We summarize the principal existing modeling approaches, show how these approaches are special cases of the proposed model, and discuss possible extensions to model more complex data structures. We present simulation and analytical results that clarify the interplay between estimated teacher effects and repeated outcomes on students over time. We also explore the potential impact of model misspecifications, including missing student covariates and assumptions about the accumulation of teacher effects over time, on key inferences made from the models. We conclude that mixed models that account for student correlation over time are reasonably robust to such misspecifications when all the schools in the sample serve similar student populations. However, student characteristics are likely to confound estimated teacher effects when schools serve distinctly different populations.

Keywords: *accountability, model misspecification, omitted-variables*

A currently active and central education policy initiative involves the use of scores on standardized achievement tests to hold educators accountable for student outcomes. This practice is a key component of most existing state accountability

This research was supported by Grant B7230 from the Carnegie Corporation of New York. The statements made and views expressed are solely the responsibility of the authors.

systems and is a cornerstone of the recently adopted federal education legislation, the No Child Left Behind Act of 2001. Most state testing programs include tests administered in a variety of subjects and across multiple grades. One of the challenges for those responsible for designing and implementing these systems is the need to combine test-score information into a single measure that provides evidence of school or teacher effectiveness. Although many states and districts rely on fairly simple score averages or differences, a few are exploring the use of more complex models that use longitudinal data on students to determine the "value added" by a particular teacher or school (Meyer, 1997). This modeling approach has taken a number of forms and is generally referred to as "value-added modeling" (VAM).

Enthusiasm for this approach stems in large part from the belief that it can remove the effects of factors not under the control of the school, such as prior performance and socioeconomic status, and thereby provides a more accurate indicator of school or teacher effectiveness than is possible when these factors are not controlled. Applications of VAM in a few jurisdictions, including Tennessee (Sanders, Saxton, & Horn, 1997) and Dallas (Webster & Mendro, 1997), have attracted the interest of large numbers of researchers and analysts; and enthusiasm for applying these methods has grown rapidly among policymakers in recent years. Of particular interest in these applications are the evaluation individual teachers and the results suggesting that teachers have large and differential effects on student learning. This article focuses on VAM estimation of teacher effects.

Despite this enthusiasm, VAM approaches have not yet been widely adopted in formal state or district accountability systems in part because VAM requires extensive computing resources and high-quality longitudinal data that many states and districts currently do not have. As data systems and computing resources improve, and as access to the necessary software for performing these analyses increases, we expect VAM to play a more substantial role in formal accountability systems.

The longitudinal student outcomes data used in VAM present many challenges for statistical modeling of teacher effects and the variability among teachers. Two are of primary interest: multiple measures on the same student and multiple teachers instructing each student. Models must account for and use the likely positive correlation among multiple measures on the same student. Class groupings of students change annually, and students are taught by a different teacher each year. The existence of teacher effects and shared but unobserved environmental variables for students within classes will contribute to positive intra-class correlation among outcomes for students in the same class. However, changing classes and teachers across years means that student outcomes do not follow the traditional nested designs of hierarchical models (Raudenbush & Bryk, 2002; Goldstein, 1995), and alternative model formulations are necessary.

There exist several models for estimating teacher effects that capture these complexities of student outcomes. To date there have been limited comparisons of

these methods. Rowan, Correnti and Miller (2002) conducted an empirical comparison of estimated teacher effects from three alternative models for longitudinal student test score data: covariate adjustment with current scores regressed on prior scores and student home and background variables; one year gains (i.e., current year score less prior year score) with adjustment for background variables; and a complex cross-classified random effects model (Raudenbush & Bryk, 2002). Details on these models are provided in the Alternative Value-Added Models Section.

The authors report that across multiple subjects and cohorts, teachers' contribution to total variability in scores (gains) ranged from 4 to 16 percent for covariate adjustment models, from 3 to 10 percent for gain score models, and from 10 to 20 percent for cross-classified models. Rowan et al. argue in favor of the cross-classified model because, unlike the other models, it decomposes within-classroom variability in growth into its systematic and residual components. While Rowan and colleagues provide an important first exploration of the robustness of teacher effects to the alternative model specifications, their study is limited in that it does not provide detailed comparison of model assumptions or likely results of violation of those assumptions. In particular, they do not investigate whether VAM really does remove the effects of factors such as prior performance and socioeconomic status, and thereby provide a more accurate indicator of teacher effectiveness than is possible when these factors are not controlled. We extend their results in several important ways.

Our primary purpose in this article is to develop and evaluate a multivariate, longitudinal mixed-model that respects the crossed grouping structures inherent to longitudinal student data linked to teachers. We also describe alternative models used for VAM of teacher effects and demonstrate that these models are special cases of the general model. We use this unifying framework to contrast the alternative approaches. We also study in detail model misspecification and its likely effects on estimated teacher effects in terms of likely systematic errors in the estimates. We use an empirical example to explore the efficiency of estimated teacher effects and consider efficiency of estimates more broadly in the Discussion Section.

A General Model for Longitudinal Student Outcomes

We start with a general model for student outcomes. Although many types of outcomes might be considered (e.g., grade completion or retention, attendance, disciplinary actions), most value-added modeling to date has focused exclusively on scores from standardized assessments. Therefore we will use the terms outcomes and test scores interchangeably.

We first present models for students from a single school system, such as a school district, state, or intermediate aggregation of districts. We also focus on scores for a single subject such as math or reading and a single cohort of students. Extensions to the models to allow for multiple school systems, subjects, or cohorts are presented after the initial model. We limit presentation to four, contiguous grades of test scores, although all results generalize to more grades.

The score data are y_{ig} for the student i's score in grade g. For notational convenience, we let $g = 0$ for the first grade of data collection, $g = 1$ for the second and so on. The model for grade 0 scores is:

$$y_{i0} = \mu_0 + \beta_0' x_i + \gamma_{00}' z_{i0} + \sum_{k=1}^{M} \lambda_{i0k}\eta_{0k} + \sum_{j_0=1}^{N_0} \phi_{i0j}\theta_j + \epsilon_{i0}, \quad (1)$$

where the η_{0k} denote grade 0 school effects (i.e., deviations in school-level means from the overall system mean) for the M schools in the system and might be considered either fixed or *i.i.d.* random normal with mean zero and variance, $\sigma^2_{\eta 0}$, i.e., $\eta_{0k} \sim N(0, \sigma^2_{\eta 0})$. The λ_{i0k} measure the proportion of grade 0 schooling that school k provided to student i. If student i did not attend school k, then $\lambda_{i0k} = 0$. If the student attended only school k then $\lambda_{i0k} = 1$; otherwise it is between 0 and 1. For students who spend only part of the year in the system, the sum of the $\lambda_{\lambda i0k}$ will be less than 1. For students who complete the entire year in this system the sum of λ_{i0k} will equal 1. Likewise, the θ_{0j} are *i.i.d.* $N(0, \sigma^2_{\theta 0})$ teacher effects for the N_0 grade 0 teachers and the ϕ_{i0j} measure the proportion of grade 0 education provided to student i by teacher j. The values of λ_{i0k}'s and ϕ_{i0j}'s are observed from administrative data and not estimated.

The x_i and z_{i0} are time invariant and time varying covariates for student i. These include student-level variables such as such as gender, race, poverty level (time invariant) and special testing circumstances (time varying; e.g., accommodations given to students with disabilities). They might also include classroom level variables such as the classroom percentage of special education students. The ϵ_{i0} are *i.i.d.* $N(0, \sigma^2_{\epsilon 0})$ residual error terms. Given the fixed effects and variance components of this model and assuming school effects are treated as random, $E(y_{i0}) = \beta_0' x_i + \gamma_{00}' z_{i0}$ and $Var(y_{i0}) = \sigma^2_{\eta 0} \Sigma\lambda^2_{i0k} + \sigma^2_{\theta 0} \Sigma\phi^2_{i0j} + \sigma^2_{\epsilon 0}$. Throughout the article, unless otherwise noted expectation and variance are conditional on the observed covariates but not random teacher or school effects.

To simplify notation, let $\eta_0 = (\eta_{01}, \ldots, \eta_{0M})'$, $\lambda_{i0} = (\lambda_{i01}, \ldots, \lambda_{i0M})'$, $\theta_0 = (\theta_{01}, \ldots, \theta_{0N_0})'$, and $\phi_{i0} = (\phi_{i01}, \ldots, \phi_{i0N_0})'$ and write Equation 1 as

$$y_{i0} = \mu_{i0} + \beta_0' x_i + \gamma_{00}' z_{i0} + \lambda_{i0}' \eta_0 + \phi_{i0}' \theta_0 + \epsilon_{i0}. \quad (2)$$

Including additional grades of outcomes breaks down the traditional nesting of students into classes because across years students are not uniquely assigned to one teacher or grouped together in the same classes. We capture this mixing, sometimes referred to as cross-classification (Raudenbush & Bryk, 2002; Goldstein, 1994), by explicitly modeling the effects of prior grade teachers (and schools) on current year scores using the following model:

$$y_{i1} = \mu_1 + \beta_1' x_i + \gamma_{11}' z_{i1} + (\omega_{10}\lambda_{i0}'\eta_0 + \lambda_{i1}'\eta_1) + (\alpha_{10}\phi_{i0}'\theta_0 + \phi_{i1}'\theta_1) + \epsilon_{i1}, \quad (3)$$

$$y_{i2} = \mu_2 + \beta_2' x_i + \gamma_{22}' z_{i2} + (\omega_{20}\lambda_{i0}'\eta_0 + \omega_{21}\lambda_{i1}'\eta_1 + \lambda_{i2}'\eta_2) + (\alpha_{20}\phi_{i0}'\theta_0 + \alpha_{21}\phi_{i1}'\theta_1 + \phi_{i2}'\theta_2) + \epsilon_{i2}, \quad (4)$$

$$y_{i3} = \mu_3 + \beta_3' x_i + \gamma_{33}' z_{i3} + (\omega_{30}\lambda_{i0}'\eta_0 + \omega_{31}\lambda_{i1}'\eta_1 + \omega_{32}\lambda_{i2k}'\eta_{k2} + \lambda_{i3}'\eta_3) + (\alpha_{30}\phi_{i0}'\theta_0 + \alpha_{31}\phi_{i1}'\theta_1 + \alpha_{32}\phi_{i2}'\theta_2 + \phi_{i3}'\theta_3) + \epsilon_{i3}. \quad (5)$$

In these models the dimensions of z_{ig} can be increasing with g if, for example, the prior values of the time-varying covariates are carried forward with (possibly) different coefficients at each grade (i.e., the model includes interactions between grade and the time varying covariates).

For $g > 0$ similar distributional assumptions are made for all the random terms as they were in Equation 2. The parameters α_{10}, α_{20}, α_{21}, etc. determine how prior year teachers contribute to current year scores. If they all equal zero, then prior year teachers have no contribution to current year scores. If they all equal 1, then teacher effects persist undiminished in perpetuity, contributing equally to the current year as they did in the past. Furthermore, when all $\alpha \equiv 1$, teachers from prior years make no contributions to gains in scores. If they are less then one and decay exponentially with the gap between test administrations, e.g., $\alpha_{gg'} = \alpha^{g-g'}$ for $\alpha < 1$, then over time the contributions of a teacher effect to students' scores follows a stationary autoregressive structure. When all $\alpha < 1$, prior year teachers contribute inversely to gains because students "regress" to the mean after leaving the teacher's class. Thus, a positive teacher effect contributes negatively to gains and a negative teacher effect contributes positively to gains. When all $\alpha > 1$, positive teacher effects have positive effects on gains and visa-versa for negative effects. The ω's function similarly for school effects.

Because each student has many unique characteristics, and abilities cannot be completely measured and incorporated into any model, scores from the same student are likely to be correlated, even after accounting for measured attributes through covariate adjustment. Therefore, we employ an unrestricted covariance matrix for the residual error terms, i.e., $Corr(\epsilon_{ig}, \epsilon_{ig'}) = \rho_{gg'}$ and variance free to change across years. Also, we employ a model where teacher and school effects are assumed to be independent across years. Alternative assumptions regarding teachers and schools are discussed below.

Together these assumptions yield that:

$$E(y_{ig}) = \mu_g + \beta_g' x_i + \gamma_{ig}' z_{ig}, \quad (6)$$

and

$$Var(y_{ig}) = \sum_{t=0}^{g} \omega_{gt}^2 \lambda_{ig}' \lambda_{ig} \sigma_{ng}^2 + \sum_{t=0}^{g} \alpha_{gt}^2 \phi_{ig}' \phi_{ig} \sigma_{\theta g}^2 + \sigma_{\epsilon g}^2, \quad (7)$$

where $\omega_{gg} = \alpha_{gg} = 1$.

Extensions to and Comments on the General Model

Although our model explicitly accounts for many important features of longitudinal score data, several details of and extensions to it merit consideration.

Teacher Effects

The statistical model presented above characterizes teacher effects as random variables that contribute to test scores. However, the teacher effects of interest are causal contributions of teachers to student achievement. The relationship between the statistical model and the causal effect depends on numerous assumptions. Implicit in our model is that the teacher has a constant effect on all students relative to other teachers in the system. Given that teacher effects might not be constant, the effect is an approximation to the teacher's average effect on students in the population that are likely to be in his or her class–assuming the model is otherwise correctly specified. In the first year of testing the model is unlikely to be correctly specified because it does not fully account for the student's history prior to this grade of testing. In this case, the estimated teacher effect will tend to include historical factors such as student background and previous educational experience that cluster by classroom. Thus, estimated teacher effects when $g = 0$ should be interpreted cautiously. Similar issues hold for school effects.

We consider two measures of teacher effects: estimates of individual teacher effects and the overall contributions of teachers to variability in student outcomes. The best linear unbiased predictor, BLUP, provides the estimate of each individual teacher's effect for mixed models such as the general models and most of the models discussed below (Searle, Casella, & McCulloch, 1992). The variance components for teacher effects ($\sigma^2_{\theta g}$s) and their ratios to the overall variability in outcomes describe the teachers' contribution to total variance.

Multiple School Systems

Implicit in the current model is the assumption that within the school system, teacher, student and possibly school effects are exchangeable, so that they can be modeled as random variables from a single teacher, student or school distribution. When pooling data from multiple systems the teachers, students and schools must remain exchangeable. If they are not, the model must be adapted so that conditional on the fixed effects and variance components of the model, the random effects can be considered exchangeable. Therefore, if we want to include multiple systems, then all fixed effects in the model should be considered system specific.

Multiple Subjects Per Grade

Often students are tested on multiple subjects such as math, reading, science, etc., at each grade. Several additional terms are necessary to extend the general model to the joint distribution of multiple subject scores per grade. The model must include separate teacher and school effects for each subject and each grade and describe how these affect all outcomes and persist over time. The model must also

specify the correlation between teacher effects, school effects and residual error terms for different subjects within and across grades.

Multiple Cohorts

Data from multiple cohorts provide repeated measures of some teacher and school effects. Therefore, we might consider jointly modeling the data from multiple cohorts with correlation in teacher and school effects across cohorts as a means of improving the statistical properties of estimated teacher and school effects. Ideally, model parameters and teacher and school effects would change across cohorts so that school and teacher improvements could be modeled and tested. However, given the reality of data availability and computational capacity, model parameters might be held constant across cohorts to improve the precision of estimates, at the possible cost of adding bias.

Multiple subjects and multiple cohorts can greatly increase the computational burden of fitting the models. Estimation requires inverting a sparse square matrix with rows for every student's scores by year and subject. The matrix grows with the square of additional subjects and cohorts. In addition, selecting the cohorts to include can be difficult. Ballou, Sanders and Wright (2003) provide a detailed discussion of this choice in one application.

Random Slopes

For simplicity of presentation, we include only fixed parameters for the covariates. The model could include school level (random or fixed) parameters and possibly teacher level random slopes for the covariates. Inclusion of teacher random slopes would provide a means of modeling teacher effects that vary across students–the teacher's effect for a student with covariates x_i and teacher j is $\theta_j + \beta_j' x_i$. However, including random slopes requires that we model the effects of these on future scores resulting in extremely complex models with demanding data requirements for obtaining precise estimates. These data demands might be unattainable in many application and simpler models might be required. In addition, including random slopes in the model obscures salient features of the model, so we do not present these models.

Measurement Error

The general model implicitly includes measurement errors in test scores as a component of the residual error terms. However, the variability in measurement error typically is not constant across the range of "true scores," where the true score is a student's level of achievement for the skills measured by the test. Therefore, the model should be extended to allow for heteroscedastic residual errors where the variance depends on the true score which in turn depends on teacher and school effects as well as student covariates and fixed effects. The variability of measurement error might be available from test publisher or explored empirically. However, because no VAM models currently considered in the literature explicitly model this nonconstant variance, we will not explicitly include heteroscedastic measurement error in our presentation.

Unequal Inter-Testing Intervals

We assume that for all students the intervals between tests is nearly constant and treat time of test as year or grade. In some settings testing intervals may vary across students, for example, when different school systems are combined for analysis. This situation makes use of the unrestricted covariance problematic and a parametric specification of the covariance matrix might be used. For example, constant correlation among scores (compound symmetry) or correlation that decays exponentially with the gap between test administrations (stationary autoregressive) or possibly a combination of these two structures might be considered.

Different Scales Across Grades

The tests used in some school systems are not designed to produce a single developmental scale across grades. For example, a school district might use tests that are not vertically linked and provide only grade specific normal curve equivalents. The general model, as specified earlier, employs an unspecified covariance matrix for residual errors that can accommodate the nonconstant variances and covariances that likely result from different developmental scales across grades. In addition, the estimation of αs and ωs allows for teacher and school effects to have different scales in the models for current and future scores as required by the different developmental scales. Thus, the model of Equations 1 to 5 is sufficiently flexible to model longitudinal data even when the developmental scales are not constant across grades provided the scales are linearly related.

Alternative Value-Added Models

Our model was motivated by considerations of the nature of longitudinal test score data. We describe alternative models that are currently used to model such data. We consider generic models such as covariate adjustment models, models for gain scores and cross-classified models. We also review two particular models used for VAM: the cross-classified model of Rowan, Correnti and Miller (2002) and the layered model of the Tennessee Value-Added Assessment System. We include both generic and specific models so that we can later draw the appropriate relationships among models and improve interpretation of results from different models. We start with the covariate adjustment and gain score models because they can be fit using standard hierarchical models software and are most widely used for modeling scores. The other models require specialized software and have achieved less widespread use.

Covariate Adjustment Models

One common approach to modeling longitudinal data is to use prior scores as covariates in models for current outcomes (Rowan, Correnti, & Miller, 2002; Diggle, Liang, & Zeger, 1996; Meyer, 1997). This is also a model commonly used in much

of the economics production function literature (Hanushek, 1992). For example, Rowan et al. consider the following model:

$$y_{ig} = \mu_g^A + \beta_g^{A\prime}\mathbf{x}_i + \gamma^{A*}y_{ig-1} + \gamma_{ig}^{A\prime}\mathbf{z}_{ig} + \lambda_{ig}'\eta_g^A + \phi_{ig}'\theta_g^A + \epsilon_{ig}^A, \quad (8)$$

with the assumptions that for each grade the school and teacher effects and the residual error terms are respectively *i.i.d.* normal random variable with mean zero, variance $\sigma_{\eta g}^{A2}$, $\sigma_{\theta g}^{A2}$ and $\sigma_{\epsilon g}^{A2}$, and independent of each other. Note that we use superscripts to distinguish the parameters from different models. Teacher and school effects also include superscripts because the different model specifications implicitly change the interpretation of these effects. Such models are often used with only two years of data and provide only one year of teacher effect estimates. However, with multiple years of testing the models typically assume that all cross-year correlation is explained by the inclusion of the prior year scores as a covariate so that $Corr(\epsilon_{ig}^A, \epsilon_{ig'}^A) = 0$ for $g \neq g'$ and prior year teacher effects do not explicitly enter the model. The complexity of cross-classification is assumed to be completely modeled by the inclusion of the covariate and the models can be specified as a traditional hierarchical model (Raudenbush & Bryk, 2002; Goldstein, 1995). Rowan et al. follow this approach, but they treat the variance components as constant and pool the data across years for estimation.

If the model is extended to allow for correlation among the residual errors across years, then standard mixed model estimation would yield biased estimates of fixed effects because of the correlation between the covariate and the residual error term. Alternative approaches that are similar to fitting the general model would be required for accurate estimation. Thus, to keep this model distinct from other alternatives, we consider only the case where residual error terms are assumed to be independent across years.

Repeated Cross-Section Models of Gains

When all scores are on the same scale, scores from adjacent grades can be differenced to obtain "gains" that are then modeled (Rowan et al., 2002; Shkolnik, Hikawa, Suttorp, Lockwood, Stecher, & Bohrnstedt, 2002). We let $d_{i0} = y_{i0}$ and $d_{ig} = y_{ig} - y_{ig-1}$ for $g \geq 1$. The model for grade $g \geq 1$ gains is

$$d_{ig} = \delta_g^G + \beta_g^{G\prime}\mathbf{x}_i + \gamma_{ig}^{G\prime}\mathbf{z}_{ig} + \lambda_{ig}'\eta_g^G + \phi_{ig}'\theta_g^G + \epsilon_{ig}^G. \quad (9)$$

The coefficient δ_g^G denotes the mean gain in grade g. Random teacher and school effects and student residual error terms follow the same assumptions as the previous model. In particular, as with the previous model the correlation across grades in ε_{ig}^G's typically is ignored and will be for our discussion. Thus, differencing scores is assumed sufficient to capture all the important cross-grade correlation structure and the complexity of cross-classification is removed by this assumption.

Some analysts (Rowan et al., 2002) include y_{ig-1} on the right hand side of Model 9. However, doing so makes Model 9 equivalent to Model 8 with γ^* replaced by $\gamma^* - 1$ (Wertz & Linn, 1970). Thus models for gain scores should not include the prior score as a covariate.

Cross-Classified Models

In their hierarchical models book, Raudenbush and Bryk (2002) (RB) develop a cross-classified model that explicitly models the cross-grade correlations and the effects of the multiple years of teachers on student outcomes. RB consider random linear growth trajectories for students. The cross-classified (CC) model for scores y_{ig} for student i in grades 0 to 3 is:

$$
\begin{aligned}
y_{i0} &= \mu^C + \mu_i + \phi'_{i0}\theta_0^C + \epsilon_{i0}^C \\
y_{i1} &= \mu^C + \gamma^C + \mu_i + \gamma_i + \phi'_{i0}\theta_0^C + \phi'_{i1}\theta_1^C + \epsilon_{i1}^C \\
y_{i2} &= \mu^C + 2\gamma^C + \mu_i + 2\gamma_i + \phi'_{i0}\theta_0^C + \phi'_{i1}\theta_1^C + \phi'_{i2}\theta_2^C + \epsilon_{i2}^C \\
y_{i3} &= \mu^C + 3\gamma^C + \mu_i + 3\gamma_i + \phi'_{i0}\theta_0^C + \phi'_{i1}\theta_1^C + \phi'_{i2}\theta_2^C + \phi'_{i3}\theta_3^C + \epsilon_{i3}^C. \qquad (10)
\end{aligned}
$$

The ϵ^Cs are assumed to be *i.i.d.* normally distributed random variables with mean zero and variance σ_ϵ^{C2}. The θ are again teacher or classroom effects and are assumed to be independently, normally distributed with constant variance across years. Each student's growth over grades is modeled with a linear trend $\mu^C + \gamma^C g + \mu_i + \gamma_i g$ and the random intercepts and slopes are assumed normally distributed with mean zero and variance τ_{00}^2, τ_{11}^2 and covariance τ_{01}. Equation 10 assumes testing is at the same regular intervals for all students, which is likely to be a reasonable approximation for many testing programs. However, if the timing of tests is available in the data and it varies appreciably across students, grade can be replaced by time since baseline testing. When time between tests is not constant, the variances and covariances for the residual error terms will vary among student as a function of variation in time between tests.

Rowan, Correnti and Miller (2002) (RCM) use a variant of the cross-classified (CC) model to specifically estimate variability of teacher effects, to which they also refer as classroom level variability. The model includes time-varying covariates for participation in educational programs (e.g., special education or gifted and talented) and age. Their model also includes time-invariant covariates for student ethnicity, family structure and socioeconomic status (SES). The time-constant characteristics are not interacted with grade and so do not contribute to the model for gains.[1] This model includes linear and quadratic terms for months since the first test with random slopes for students on the linear term. RCM also include random school effects in their model.

Tennessee Value Added Assessment System, Layered Model

The Tennessee Value-Added Assessment System (TVAAS) produces estimated teacher effects using a model that William Sanders and colleagues, (Sanders et al.,

1997), call the layered model because the model for later years adds layers to the model for earlier years. For a single subject and cohort of students from one school system, the layered model is:

$$\begin{aligned}
y_{i0} &= \mu_0^T + \phi'_{i0}\theta_0^T + \epsilon_{i0}^T \\
y_{i1} &= \mu_1^T + \phi'_{i0}\theta_0^T + \phi'_{i1}\theta_1^T + \epsilon_{i1}^T \\
y_{i2} &= \mu_2^T + \phi'_{i0}\theta_0^T + \phi'_{i1}\theta_1^T + \phi'_{i2}\theta_{i2}^T + \epsilon_{i2}^T \\
y_{i3} &= \mu_3^T + \phi'_{i0}\theta_0^T + \phi'_{i1}\theta_1^T + \phi'_{i2}\theta_{i2}^T + \phi'_{i3}\theta_{i3}^T + \epsilon_{i3}^T.
\end{aligned} \tag{11}$$

The ϵ_{ig}^Ts are assumed normally distributed and independent across students. Within a student the variance-covariance matrix of the ϵs is unrestricted allowing for different variance at each time point and possibly nonzero and nonconstant correlation of scores from different years (grades). The variance-covariance parameters are assumed constant across all students.

As with the other models we consider, the teacher effects are assumed to be independent normally distributed with zero mean. Effects are assumed to be independent both within and across years. TVAAS allows the variance of teacher effects to vary across grades.

Equation 11 simplifies the TVAAS model in several ways. First, TVAAS often models data from multiple school systems. When doing so, Model 11 is extended to allow only the means to be system dependent with all other coefficients held constant across systems. TVAAS uses data from grades 2 through 8 and considers multiple subject tests per grade. The TVAAS model allows correlation between scores from the same student across subjects (and grades). However, TVAAS assumes that teachers have separate and independent effects for each subject, even if they teach multiple subjects. Thus, models for multiple subjects would be similar to models with multiple years and a single subject and our discussions do not lose generality by considering only a single subject.

A Comparison of Alternatives and the General Model

The model given in Equations 1 to 5 is sufficiently general to include all the alternative models as special cases. Figure 1, which summarizes the relationships between all the models, shows the generality of our new model by including it at the center with all models pointing to it. This section provides details on the relationship of each alternative to the general model and on the relationships among models. In particular, as shown in Figure 1, we demonstrate that the gain score model and CC model, when time is constant across students, are special cases of the layered model. For comparisons, we consider five features of the models: parameterization of the overall time trend; inclusion of covariates; the distribution of residual error terms; the persistence of teacher effects on future outcomes; translations between modeling scores and gains.

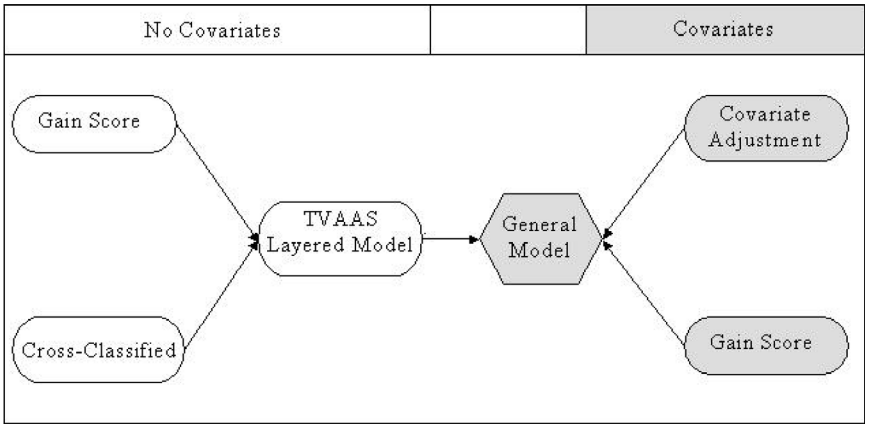

FIGURE 1. *Relationship among models. Without covariates, gain scores and the cross-classified model are special cases of the layered model with restrictions to the overall time trend and/or the distribution of residual errors. The layered model is a special case of the general model with restrictions to the* α*s and without covariates. The covariate adjustment and gain-score model with covariates are special cases of the general model with restrictions to the distribution of residual errors and the* α*s.*

Covariate Adjustment and the General Model

We expand the covariate adjustment Model 8 to obtain the alternative expression for y_{ig} under model:

$$\begin{aligned} y_{ig} = {} & \mu_g^A + \sum_{t=0}^{g-1} \gamma^{A*g-t} \mu_t^A + \left(\beta_g^A + \sum_{t=0}^{g-1} \gamma^{A*g-t} \beta_t^A \right)' \mathbf{x}_i + \gamma_g^{A'} \mathbf{z}_{ig} + \\ & \sum_{t=0}^{g-1} \gamma^{A*g-t} \gamma_t^{A'} \mathbf{z}_{it} + \lambda_{ig}' \eta_g + \sum_{t=0}^{g-1} \gamma^{A*g-t} \lambda_{it}' \eta_t + \phi_{ig}' \theta_g + \\ & \sum_{t=0}^{g-1} \gamma^{A*g-t} + \phi_{it}' \theta_t + \epsilon_{ig}^A + \sum_{t=0}^{g-1} \gamma^{A*g-t} + \epsilon_{ig}^A, \end{aligned} \tag{12}$$

which is the general model with the following restrictions:

- $\mu_g = \mu_g^A + \sum_{t=0}^{g=1} \gamma^{A*g-t} \mu_t^A$;
- $\beta_g = \beta_g^A + \sum_{t=0}^{g=1} \gamma^{A*g-t} \beta_t^A$;
- $\gamma_g = \gamma_g^A + \sum_{t=0}^{g=1} \gamma^{A*g-t} \gamma_t^A$;
- $\epsilon_g = \epsilon_{ig}^A + \sum_{t=0}^{g=1} \gamma^{A*g-t} \epsilon_{it}^A$; and
- $\alpha_{gg'} = \omega_{gg'} = \gamma^{A*g-g'}$

Thus, the covariate adjustment model restricts the residual error terms to be autoregressive. It also restricts teacher and school effect to have the same autoregressive structure.[2]

Because of measurement error in observed test scores, analysts typically consider the correct specification for the right-hand side of Model 12 to include the "true score" for the prior achievement rather than the error prone observed prior year test score. That is $y_{ig} = \mu_g^A + \beta_g^{A\prime}\mathbf{x}_i + \gamma^{A*}u_{ig-1} + \gamma_{ig}^{A\prime}\mathbf{z}_{ig} + \lambda_{ig}'\eta_g^A + \phi_{ig}'\theta_g^A + \epsilon_{ig}^A$, where u_{ig} is the true score, ξ_{ig}^A is measurement error and $y_{ig} = u_{ig} + \xi_{ig}^A$. The correspondence between the general model and the covariate adjustment model also holds when the covariate adjustment model is specified in terms of a regression on the true score rather than the measure with error. However, in this case, $\epsilon_g = \zeta_{ig}^A + \sum_{t=0}^{g-1}\gamma^{A*g-t}\zeta_{it}^A + \xi_{ig}^A$, where ζ_{ig}^A is the residual error term in the true score and $\epsilon_{ig}^A = \zeta_{ig}^A + \xi_{ig}^A$.

Gain Scores and the General Model

The gain score Model 9 uses first differences of the test scores. Under the general model, first differences for $g > 0$ can be written as:

$$d_{ig} = (\mu_g - \mu_{g-1}) + (\beta_g - \beta_{g-1})'\mathbf{x}_i + (\gamma_g - \gamma_{g-1})'\mathbf{z}_{gi} + \lambda_{ig}'\eta_g + (\omega_{g,g-1} - 1)\lambda_{ig-1}'\eta_{g-1} + \sum_{t=0}^{g-2}(\omega_{g,t} - \omega_{g-1,t})\lambda_{it-1}'\eta_t + \phi_{ig}'\theta_g + (\alpha_{g,g-1} - 1)\phi_{ig-1}'\theta_{g-1} + \sum_{t=0}^{g-2}(\alpha_{g,t} - \alpha_{g-1,t})\phi_{it}'\theta_t + \epsilon_{ig} - \epsilon_{ig-1}. \quad (13)$$

The gain score Model 9 equals the general model with the following restrictions

- all of the αs and ωs equal 1;
- ϵs restricted to $\epsilon_{ig} = \epsilon_{ig-1} + \epsilon_{ig}^G$;
- $\delta_g^G = \mu_g - \mu_{g-1}$;
- $\beta_g^G = \beta_g - \beta_{g-1}$; and
- $\gamma_g^G = \gamma_g - \gamma_{g-1}$.

Thus the gain score model implicitly assumes that teacher (school) effects persist undiminished into the future and that residual errors follow an autoregressive unit root process with no unobserved student heterogeneity in gain scores, i.e., zero correlation across time in the residual error terms for a student's gains.

Covariate Adjustment and Gain Score Models

In general there is considerable debate over whether to model gains or fit covariate adjustment models. Many authors argue against the covariate adjustment model (Rowan et al., 2002; Thum, 2002). They argue that the covariate adjustment model is not a model of growth and that because of measurement error, estimates from Model 8 are inconsistent. However, other authors note that gains are more variable

and less "reliable" than scores and that gains can produce biased estimates of the parameters of structural or causal models. For modeling longitudinal data and estimating teacher effects, both models make restrictive assumptions about error terms and persistence of teacher effects. Our general model captures the features of both the gains and covariate adjustment models and can test the assumptions of each model yielding an empirical determination of the best model. We have not studied the relative efficiency of the two models in the context of teacher effect estimation. However, for estimating group means the covariate model produces more efficient estimates, when the model is correct (Feldt, 1958). Yang and Tsiatis (2001) consider relative efficiency in a semiparametric model assuming only the first two moments of the joint distribution. They find that change score modeling tends to be more efficient than covariate adjustment for estimating group means in an experimental setting.

CC, TVAAS and the General Model

Compared to the general model of Equations 1 to 5, the CC model with constant testing times places restrictions on both fixed effects and the distribution of the residual error terms. The CC model uses a linear trend in time rather than allowing separate means for each time point, $\mu_g = \mu^C + \gamma^C g$. In addition, the CC model forces the βs and γs to be identically zero and, like the gain score model, assumes that all the αs and ωs equal 1.

CC also restricts the variances and covariances of multiple observations from a student. CC assumes that $\epsilon_{ig} = \mu_i + \gamma_i g + \epsilon^C_{ig}$ so that $Var_{CC}(\epsilon_{ig}) = \tau^2_{00} + 2g\tau_{01} + g^2\tau^2_{11}$ and $Cov_{CC}(\epsilon_{ig}, \epsilon_{ig'}) = \tau^2_{00} + (g + g')\tau_{01} + gg'\tau^2_{11}$ and the variability of scores necessarily increases over time for sufficiently large values of g. In the general model, error term variances and covariances are unrestricted.

The layered model is also a special case of the general model that excludes school effects and covariates and where the α's are assumed to equal one. However, unlike the CC model, the TVAAS layered model places no restrictions on the overall grade specific means (the μ^T_gs) and places no restrictions on the covariance matrix of the repeated test scores from a student. Thus, CC is a special case of both the general model and the TVAAS model with restrictions to the growth over time in the overall means and restrictions to the distribution of residual error terms.

CC, TVAAS and Gain Scores

When all of the αs equal 1, as in both the TVAAS layered model and the CC cross-classified model, first differences depend only on the current year teachers. Taking first differences of adjacent year scores, yields that for the CC model

$$d_{ig} = \delta^C_g + \phi'_{ig}\theta^C_g + \zeta^C_{ig}, \tag{14}$$

where $\delta_0 = \mu$, $\zeta_{i0} = \mu_i + \epsilon^C_{i0}$ and $\delta_g \equiv \gamma^C$ and $\zeta_{ig} = \gamma_i + \epsilon^C_{ig} - \epsilon^C_{ig-1}$ for $g = 1, 2, 3$. Assuming all student complete all the tests so that raw scores can be transformed to gain scores, the CC model is a multi-grade gain score model, where the mean gain is

assumed constant across grades and variance-covariance matrix for residual error terms from the same student is not diagonal (i.e., gains are not independent across grades) but given by:

$$\mathbf{R}_1 = \begin{pmatrix} \tau_{00}^2 & \tau_{01} - \sigma_\xi^2 & \tau_{01} & \tau_{01} \\ \tau_{01} - \sigma_\xi^2 & \tau_{11}^2 - 2\sigma_\xi^2 & \tau_{11}^2 - \sigma_\xi^2 & \tau_{11}^2 \\ \tau_{01} & \tau_{11} - \sigma_\xi^2 & \tau_{11}^2 - 2\sigma_\xi^2 & \tau_{11}^2 - \sigma_\xi^2 \\ \tau_{01} & \tau_{11}^2 & \tau_{11}^2 - \sigma_\xi^2 & \tau_{11}^2 - 2\sigma_\xi^2 \end{pmatrix}. \quad (15)$$

Similarly the layered model for differences is:

$$d_{ig} = \delta_g^T + \phi'_{ig}\theta_g^T + \zeta_{ig}^T, \quad (16)$$

where $\delta_0 = \mu_0^T$, $\zeta_{i0} = \epsilon_{i0}^T$ and $\delta_g = \mu_g^T - \mu_{g-1}^T$, $\zeta_{ig} = \epsilon_{ig}^T - \epsilon_{ig-1}^T$ for $g = 1, 2, 3$ and the residual variance-covariance matrix remains unrestricted.

When all students have complete data, the d_gs uniquely map to the y_gs and d_gs are sufficient for the y_gs and teacher effects and variance components can be estimated fully efficiently from modeling either the gains or the raw scores. Thus, both the CC and the layered models essentially determine teacher effects by gains for $g > 0$, although the inclusion of multiple years of gains and the intra-student correlation result in adjusted gains that account for student performance across all years.

The CC and layered models use data from all students, even those with partially complete records. Gain score modeling uses only students with both years of data unless imputation or another missing data method is applied. Again the CC and layered models are extensions of gain score modeling. Loss of incomplete records can not only decrease precision, but also introduces bias when the distribution of missing test scores are not missing completely at random (Little & Rubin, 1987). CC and layered models are robust to missing data provided data are missing at random (Little & Rubin, 1987).

When the data are incomplete the simple translation from raw to gain scores is not possible. However, because in both the CC and layered models gains link to only one teacher the estimation procedure essentially involves two steps: implicit imputation of values for unobserved gains using the observed scores; followed by estimation of teacher effect using the means of both the imputed and observe gains.

In the general model we introduced, gain scores depend on multiple teachers. Thus, estimated teacher effects will depend on more than adjusted gains as they do in part for incomplete data with the CC and layered models. For correctly specified models this distinction might be of little consequence. However, with omitted covariates, the layered and cross-classified models will tend to be most sensitive to omitted covariates that effect gains, while for the general model covariates that effect levels might also confound estimated effects.

Implications of Model Restrictions

The general model demonstrates that adequately capturing the complexities of longitudinal student test score data requires an extremely complex model. As described above, models previously proposed or used for VAM have tended to place restrictions on the α, β and γ parameters as well as on the distribution of the residual error terms. In this section we consider the possible ramifications of those restrictions.

Restrictions to the Residual Error Distribution

If the model has correctly specified the overall time trend and includes all relevant covariates, then restrictions on the residual error structure even if incorrect should not bias estimated coefficients (Diggle et al., 1996). However, misspecified residual error distributions can result in inefficient estimates and possibly increase the systematic errors in estimated teacher effects. For example, in a small empirical study we found that ignoring the correlation in gain scores across years resulted in a substantial loss of efficiency in the BLUP estimates of teacher effects. More generally, although parametric specification of the covariance matrix might improve the efficiency of estimates by reducing the number of estimated coefficients, these models might increase the overall error in the BLUPS, if the parametric model is poorly specified. Most importantly, if the parametric model is poorly specified the resulting errors in estimated teacher effects could be systematically related student characteristics.

The cross-sectional gains and covariate adjustment models assume no correlation in the residual error terms across years and are traditional hierarchical models with students nested within one teacher/class. In such traditional hierarchical models the estimate of the teacher variance component depends on the difference in the within and between teacher (or class) mean sums of squared errors in the residual score or gain after adjusting for the included covariates (Searle et al., 1992; Snijders & Bosker, 1994). The estimate of the variance in the teacher effects depends on the classroom level variance in the residual scores or gains. In addition, provided students have only one teacher each year, standard software can be used to estimate the parameters of these models and obtain estimated teacher effects.

Cross-classified and layered models for differences, d_{ig}, take on the familiar look of a hierarchical model. At each grade students' gain scores are nested in one class and the model involves only one teacher effect which will depend on the gains. When pooled across grades the intra-student correlation in scores as specified by the $\mathbf{R}_1$ matrix (Equation 15), or its alternative from the layered model, results in the ensemble of gains across grades all contributing to every estimated teacher effect because a student's gains in any one year are adjusted to account for his or her gains in other years. As one might expect, using information from multiple years yields more efficient estimates of teacher effects than the cross-sectional gains model.

Increasing Variability over Grades

When $g > -2\tau_{01}/\tau_{11}^2$ (e.g., whenever $\tau_{01} > 0$), the CC model induces increasing variance with grade. Furthermore both the cross-classified and the layered model

assume that variability due to teacher increases with grade. However, the variability of scaled scores on standardized tests do not necessarily increase with grade. For example, the standard deviation of scores in the Prospects data used by RCM remains at about 50 for every grade for the grade 1 cohort. [See (Rowan et al., 2002) and (Puma, Karweit, Price, Ricciuti, Thompson, & Vaden-Kiernan, 1997) for details on the Prospects data.] More generally there is considerable debate in the measurement community on the appropriate behavior of variability over time on true developmental scales (Burket, 1984; Hoover, 1984a, 1984b; Yen, 1986). Thus, the fit of the CC or layered models to some data sets will be poor. We cannot predict the consequences of lack of fit on parameter estimates, but in the example given below the layered model presents a very different picture of teacher effects across grades than the general model, which does not require variability due to teachers to increase with grades.

Omitted Covariates

The layered model and the CC model as presented include no covariates. Even when models include student characteristics, the administrative test score data available for most value-added modeling tend to include only limited information on student characteristics. Thus, omitted covariates are a possible problem for any VAM application.

The inclusion of intra-student correlation of scores complicates the assessment of the effects of omitted covariates in value-added models. Some analysts have suggested that the inclusion of intra-student correlation essentially removes the effects of omitted covariates. As we will show, this is not true in general. The impact of omitted covariates on estimated teacher effects in the presence of intra-student correlation is subtle depending on both the distribution of the omitted covariates and the assignment of students to teachers. We consider three different scenarios for the distribution of omitted covariates and the assignment of students to teachers, and examine the impact on teacher effects in each case.

Randomly Distributed Omitted Variables

To ground our discussion we will use the layered model as the basis for considering the effects of omitted covariates. Because other models tend to be similar, the results presented here will apply to other models as well. We assume that the Model 11 is correct except that a covariate u_i is omitted. Without loss of generality we assume that u_i is a scalar rather than a vector random variable for each student and we assume it is uncorrelated with all the other terms in the model.

In this scenario we assume that the omitted covariate is randomly distributed and the intra-class correlation of u_i is zero. Because u_is are not correlated within classes and are uncorrelated with the teacher effects or other terms in the model, the covariate is just another component of residual error. Even if u_i are constant over time (as the subscripting implies) and the omitted covariate contributes to intra-student correlation in scores, this should not bias results because the model accounts for that correlation. Thus, this type of omitted covariate should have

little effect on the estimated teacher effects and the estimate of the teacher variance component.

Omitted Variables Cluster by Class

For the general setting wherein the data are incomplete, the algebra for determining the effects of omitted covariates produces intractable analytic forms. Therefore we consider the case where the data are complete and we can use gain scores to specify either the layered or cross-classified models. Models fit to data with partially complete student records are unlikely to result in more complete removal of confounding effects of omitted covariates. Thus, the results presented in this section are unlikely to overstate the possibility of problems resulting from omitted covariates. For simplicity we consider cases where students are in only one class each year, so that ϕ_{ig} equals a vector of N_g zeros except for the row corresponding to the student's teacher in grade g, which equals 1. We add to the model described above the assumption that $Cov(u_i, u_{i'} | \phi_{ig} = \phi_{i'g}) = \rho\sigma_x^2$. That is, the distribution of the omitted variable is heterogeneous across classes. In the canonical example, which we are almost always considering, the mean value of the omitted variable varies across classes. If we ignore intra-student correlation then estimated teacher effects include this omitted covariate. To see this, let $\hat{\theta}$ denote the vector of estimated teacher effects. These estimates solve the mixed model equations (Searle et al., 1992)

$$\hat{\theta} = (\Phi'\Phi + \mathbf{D}^{-1})^{-1}\Phi\mathbf{r}, \tag{17}$$

where $\mathbf{\Phi}$ is a matrix with rows equal to ϕ'_{ig} and sorted by student and grade, $\mathbf{D}$ is the variance covariance matrix for the teacher effects (a diagonal matrix by assumption) and $\mathbf{r}$ is the vector of residuals that result from subtracting the estimated mean (based on fixed effects) from the vector of test scores. The matrix $\mathbf{\Phi}'\mathbf{\Phi}$ is diagonal and $\mathbf{\Phi}'\mathbf{r}$ is a vector of classroom means. Therefore, each teacher effect depends only on the mean of the class's residuals which includes the mean of the omitted covariates.

When the model includes intra-student correlation the mixed model equations are

$$\hat{\theta} = (\Phi'\mathbf{R}^{-1}\Phi + \mathbf{D}^{-1})^{-1}\Phi\mathbf{R}^{-1}\mathbf{r}, \tag{18}$$

and the first matrix is not diagonal. Thus estimated teacher effects mix residuals across classrooms and, at least partially, undo the confounding of the effects of teachers and the means of the omitted covariates. The exact amount of reduction in confounding depends on $\mathbf{\Phi}'\mathbf{R}^{-1}\mathbf{\Phi}$ which, in turn, depends on the mixing of students between teachers across years and the nature of the intra-student correlation.

We present the following heuristic argument to provide insight into how cross-grade correlation mitigates the confounding of omitted covariates and true teacher effects. Greater details can be found in Appendix 1. We restrict attention to each

student's sample of gain scores. Centering each student's gains around his or her average gain (i.e., including student "fixed effects" in the model) removes any systematic student level effects on gains. The CC and TVAAS models do not completely center each student's gains around his or her mean gain. Rather, student gains are centered around the empirical Bayes estimate of the student's mean. The empirical Bayes estimate will "shrink" the student's mean back toward the average for all students *and* it will adjust for the empirical Bayes estimates of the teacher effects. The estimation procedure can be thought of as iteratively estimating teacher effects based on adjusted gains and then readjusting student gains based on the updated teacher effects. As a result of this process, empirical Bayes estimates of student means are adjusted back toward classroom means after adjusting for the individual students means. This implies that more of the between classroom variance in gains will remain in the final estimate of the variability of the teacher effects than if the students' raw mean had been used to adjust his or her gains. Adjusting for the student mean would remove all the student specific effect not under the control of the school. However, our heuristic discussion implies that modeling the correlation in scores across grades neither adjusts for the students' mean nor necessarily removes all the student specific effects. In particular, heterogeneously grouped omitted covariates that predict gains can contribute to estimated teacher effects in the CC and layered models.

Omitted variables differ by strata

Suppose that the population of teachers can be grouped into strata where the teachers within a stratum teach classes with some overlapping students, but students do not overlap across strata. For example, suppose the population contains two schools and no students switch schools. While there is no overlap across the schools, there is overlap within schools, because the grade $(g + 1)$ classes contain students taught by one of the grade g teachers in the prior year.

Now suppose that the mean of the omitted covariate varies across strata. The class means of the omitted covariate will vary across strata, resulting in intra-class correlation for the omitted variable when classes are pooled across strata. As discussed in the previous scenario, when intra-student correlation is ignored (e.g., in the gain model) the intra-class correlation in the omitted covariate leads to confounding or errors in the estimated teacher effects that correlate with the omitted covariate.

However, unlike the previous scenarios, modeling intra-student correlation in scores will not necessarily reduce bias in the estimated teacher effects. The mixed model equations are again given by Equation 18. However, stratification of students and teachers implies that for each stratum, $\phi_{ij} = 0$ for every student who is in the stratum and every teacher who is not. Thus, Φ contains blocks of zeros corresponding strata and the product $\mathbf{\Phi'R}^{-1}\Phi$ with the block diagonal $\mathbf{R}^{-1}$ is also block diagonal with blocks $\mathbf{A}_h$ corresponding to strata. Furthermore the elements of $\mathbf{\Phi'R}^{-1}\mathbf{r}$ corresponding to any stratum h equals $\mathbf{B}_h\mathbf{r}_h$ for a matrix $\mathbf{B}_h$ that depends only on Φ and $\mathbf{R}^{-1}$. Thus, the estimated effects for teachers from stratum h equal $\mathbf{A}_h\mathbf{B}_h\mathbf{r}_h$. Given that

the expected value of $\mathbf{r}_h$ depends on the stratum mean of the omitted covariate, the teacher effects are biased. Across the strata the errors in the estimated teacher effects are correlated with the mean of the omitted covariate.

Note that the confounding as a result of stratification in the population can be eliminated by including stratum means in the model. However, including stratum means might be undesirable in practice where the average teacher effects might also vary across strata. If such heterogeneity across strata in teachers exists then including stratum means effectively makes all inferences about teachers relative to the stratum and could result in underestimation of the variability of teacher and bias in estimated teacher effects. We believe that this is one of the most difficult issues arising from the use of VAM to estimate school or teacher effects, and we return to it in the discussion.

Similarly, the inclusion of student level covariates is not necessarily the solution to bias that results from stratification of the population. The available covariates might not include all factors that effect scores and differ across strata. In addition, if teacher effects are correlated with the characteristics of the students they teach, for example, if the most effective teachers teach in affluent schools, then including covariates in the model will tend to remove part of the teacher effect (Ballou, Sanders, & Wright, 2004; Raudenbush & Willms, 1995). Such over correcting for covariates will tend to result in systematic errors in estimated teacher effects and bias low estimated variance components. We return to this point in the discussion section.

Because the models are for gains, the bias results when mean gain scores differ across strata. Effects that are constant for students but unrelated to gains can differ across strata without resulting in bias. For the general model and partially complete data, we cannot translate to gains and the implication is that covariates that contribute to raw scores as well as those that contribute to gains might confound the estimated effects.

Simulation study results

We conducted a very small simulation study to demonstrate the mathematical results described above. We generated data for 200 students grouped into 10 classes of 20 for each of four grades. Half the students were slow gainers and half were fast. The difference between groups was roughly a gain score standard deviation unit. The teacher effects equal evenly distributed percentiles of the normal distribution with variance 1.0 at each grade and with two teachers at each designated value. Students were grouped into two schools and they did not switch schools. At each grade student gains depended upon gainer status, teacher and both student by grade and student error terms. Within each grade, the student component accounted for 70% of the total student residual error. The variance of total student residual errors was 1.0.

We repeated the simulation to match the three scenarios discussed above. For the first scenario, slow and fast gainers were randomly distributed across schools and teachers. For the second, in each school and at each grade, half the classes had

5 fast and 15 slow gainers and half had 15 fast and 5 slow. The classes were randomly assigned to teachers at each grade and every student had equal probability of being in each class. For the final scenario, one school had an expectation of 25% fast gainer and 75% slow and the other school had 75% fast and 25% slow gainers. This basic example allows for clear evaluation of the model in unrealistic settings that highlights the effects of modeling intra-student correlation.

Under the first scenario, the correlation between estimated teacher effects and the percent of fast gainers in the class was moderate (.45) when the model did not explicitly model intra-student correlation. However, when we explicitly modeled the intra-student correlation (with an unstructured variance-covariance matrix) the correlation was very low (–.04). Thus, when omitted covariates contribute to random student effects that are not correlated within classes, they can result in errors but the layered or CC model will tend to be robust to that correlation. In this case, accounting for intra-student correlation is analogous to including a random block term in designed experiments–the blocks are students, and accounting for the intra-student correlation is analogous to modeling blocks. Because not all teachers are in all blocks, we must, as is the case with this scenario, have sufficient overlap among teachers and students to estimate the teacher effects.

As predicted above, under the second scenario (i.e., heterogeneity among classes in the proportion of high gainers) models that ignored intra-student correlation yielded estimated teacher effects that were highly correlated (.79) with the means of the omitted covariate. However, modeling the intra-student correlation greatly reduced the correlation (to .47) between estimated teacher effects and the class mean of the omitted variable. Thus, modeling student correlation partially mitigates the omitted covariate bias in the BLUPS in this scenario with extensive mixing of students across classes.

The effects on the estimated variance components for teacher effects are also not obvious, but in our small simulation we found that modeling one year of gains or ignoring the intra-student correlation resulted in upward bias in the estimated teacher effect. If we control for gainer status we have no apparent teacher effect, that is, we remove 100% of the spurious variance in teacher effects. Centering each student about his or her raw mean also removed 100% of the spurious variance component. When we explicitly modeled intra-student correlation, the bias was greatly reduced and small, 88 percent of the spurious variance was removed. Thus, we again find that explicitly modeling the intra-student correlation appears to greatly reduce the biasing effects of confounding on the teacher variance component when students are heterogeneously grouped but not stratified.

When students were stratified as in the third scenario, estimated teacher effects differed across strata: estimated teacher effects were systematically too large in the stratum with a disproportionate number of fast gainers, while they were too small in the other stratum. The correlation between the class average of the omitted variables and the teacher effects was high (.79) even after modeling intra-student correlation. If one student crossed the strata (e.g., a very small amount of school transfer) the estimated teacher effects were still correlated with the average of the omitted covariate.

As transfer leads to complete mixing, the errors will eventually be only weakly correlated with the omitted covariates but the amount of mixing necessary will depend on the specific nature of the covariate, strata and class sizes and the strength of the covariate for predicting gains. However, including random student effects did greatly reduce the bias in the estimated teacher variance component.

Teacher Effects Persist with Equal Effect into the Future

Models used to date for estimating teacher effects (TVAAS and RCM) assume that these effects persist undampened into the future, i.e., all the α's are identically equal to one. The validity of this assumption has never been fully explored, and while there is evidence that some teacher effects are long-lasting (Pederson, Faucher, & Eaton, 1978), there is considerable reason to conjecture that a teacher's effect will dampen over time as students grow and are exposed to other teachers and other learning experiences. Therefore, we consider model misspecification where a model is fit treating all α's fixed at one but where the data are generated from a model with α's less than one. The misspecified model can be written as follows (for ease of notation we will consider the TVAAS model that excludes covariates and consider only grade 2 for now):

$$\begin{aligned} y_{i2} &= \mu_2 + \alpha_{21}\phi'_{i1}\theta_1 + \phi'_{i2}\theta_2 + \epsilon_{i2} \\ &= \mu_2 + \phi'_{i1}\theta_1 + \phi'_{i2}\theta_2 + (\alpha_{21} - 1)\phi'_{i1}\theta_1 + \epsilon_{i2}. \end{aligned} \tag{19}$$

Thus, model misspecification results in an omitted covariate, $([\alpha_{21} - 1]\phi'_{i1}\theta_1)$. Because students in a class typically will have been in class together in the past, the omitted covariates will be correlated among students in the same class. The means of the omitted covariate will differ by strata if teacher effects differ systematically across schools with little or no student transfer. Thus on the basis of the results in the previous section, we can expect this omitted variable to have minimal contribution to systematic errors in the teacher effects if teacher effects do not cluster and result in possible bias if teachers do cluster among schools.

We explored this bias with another small simulation. We randomly generated scores according to the general model where correlation among scores from the same student was 0.7 and where $(\alpha_{21}, \alpha_{31}, \alpha_{32}) = (0.8, 0.64, 0.8)$, $(0.5, 0.25, 0.5)$ or $(0.19, 0.11, 0.30)$ and we had 20 students in each of 20 classes per grade, four in each of five schools. We simulated data grades 2 to 5 with the grade 3 score independent of the grade 2 teacher and we modeled the gain score because our data are complete. We simulated data where the true teacher effects did not cluster and where they did with an intra-school correlation for teacher effects of 0.5 or 0.9. We fit our model only to the gains for grades 3, 4 and 5.

We used the correlation between the true teacher effects and the estimated teacher effects to measure the impact of model misspecification–low correlation indicates negative impact due to misspecification. We find that even with this misspecified model the correlation between the true teacher effects and the estimates

is high–above 0.75 when intra-school correlation for teachers equals 0.9 and above 0.8 otherwise. In fact, the correlation between estimated and true effects is nearly as large for the misspecified as the correlation between the estimate and the true effects when the estimation model is correct, i.e., the data generating model is the layered model. For example, when $(\alpha_{21}, \alpha_{31}, \alpha_{32}) = (0.8, 0.64, 0.8)$ the correlation for misspecified models is over 99% as large as for the correctly specified model. Even when $(\alpha_{21}, \alpha_{31}, \alpha_{32}) = (0.19, 0.11, 0.30)$ the correlation between estimates and true effects is at least 95% as large as the corresponding correlation for a correctly specified model, with the lowest value occurring when teachers are very highly clustered. Thus, misspecification does not appear to greatly degrade the estimated teacher effects.

Example

To explore the feasibility of fitting the model, and to estimate the persistence of teacher effects, we analyzed student achievement data from 678 students from a large suburban school district. The students in the sample attended a purposively chosen sample of five elementary schools selected from the district. The chosen schools have similar proportions of free or reduced price lunch eligible (FRL) students. While schools in the district are highly heterogeneous with FRL rates ranging from less than 10% to over 80%, the FRL rates for the sampled schools range from 11% to 17%. A representative sample of the district's schools would tend to violate the model assumptions because the percent of FRL students predicts school test score means, even above and beyond the effects of lunch status on individual student outcomes.

Using mathematics scaled scores from the Stanford 9 achievement test for third, fourth and fifth graders in the five sampled schools, we fit the general model without covariates and with lunch status as a student level covariate. For comparison we also fit the layered model–i.e., our general model with no covariates and with all α's constrained to equal one. We fit the models by maximizing their respective likelihoods via the nonlinear function optimizer in the R statistical language (Ihaka & Gentleman, 1996).

Table 1 presents the estimated parameters and standard errors for the three models. For each model the table presents the teacher and residual error variance components ($\sigma^2_{\theta 3}$,$\sigma^2_{\theta 4}$ and $\sigma^2_{\theta 5}$ and $\sigma^2_{\epsilon 3}$,$\sigma^2_{\epsilon 4}$ and $\sigma^2_{\epsilon 5}$, respectively). It also presents the cross-grade intra student correlations (ρ_{34}, ρ_{35} and ρ_{45}) and the persistence measures for teacher effects (α_{34}, α_{35} and α_{45}) which by assumption are set to 1 for the layered model. Finally the table presents the log-likelihood for each model and the likelihood ratio test statistics and p-values for comparing nested models (general without lunch status to the layered model or general with to general without lunch status).

There is little evidence for including lunch status in the general model. The likelihood ratio test for the inclusion of lunch status is 1.10 with 1 degree of freedom ($p = .294$) and the parameter estimates are extremely similar between the two models. As a result the estimated teacher effects are nearly identical for the two models (correlation ≈ 1).

TABLE 1
Parameter Estimates, Standard Errors and Log Likelihood at the MLE

	Estimate			*SE*		
Parameter	General nocov	General FRL	Layered nocov	General nocov	General FRL	Layered nocov
$\sigma^2_{\epsilon 3}$	1454	1443	1495	94	94	100
$\sigma^2_{\epsilon 4}$	1313	1306	1284	84	84	83
$\sigma^2_{\epsilon 5}$	1258	1244	1222	81	81	81
ρ_{34}	0.83	0.82	0.80	0.02	0.02	0.02
ρ_{35}	0.82	0.82	0.79	0.02	0.02	0.02
ρ_{45}	0.84	0.84	0.82	0.01	0.02	0.02
$\sigma^2_{\theta 3}$	139	137	76	71	71	51
$\sigma^2_{\theta 4}$	74	74	92	34	35	42
$\sigma^2_{\theta 5}$	51	52	33	24	25	17
α_{34}	0.2	0.2	1	0.2	0.2	
α_{35}	0.1	0.1	1	0.2	0.2	
α_{45}	0.3	0.3	1	0.2	0.2	
μ_{frl}	NA	–2.03	NA		2	
max LL	–5219.82	–5219.27	–5243.44			
LRT statistic	NA	1.10 ($p = .294$)	47.24 ($p < .001$)			

Note. general nocov = general model without covariates
general frl = with student FRL status
layered nocov = layered model without covariates

The general model fits the data substantially better than the layered model and there is strong evidence against the hypothesis that $\alpha_{34} = \alpha_{35} = \alpha_{45} = 1$. The likelihood ratio statistic is 47.24 with 3 degrees of freedom ($p < .001$). The estimated α's are all substantially less then 1 and none is significantly greater than zero suggesting that for these students, prior teacher effects contribute little to current outcomes. While results from such a small sample of schools and with such a limited set of covariates must be interpreted cautiously, they suggest that the persistence of teacher effects and our model warrant further exploration.

The general and layered models also provide notably different estimates of the teacher variance components. For grade 3, the estimated component for the general model is 83 percent larger than the corresponding estimate for the layered model (138.6 compared to 75.7). For grade 4 the estimate for the general model is only 80% as large (73.6 versus 91.6); while at grade 5, the estimate for the general model is 54% larger (51.2 versus 33.1). As a result the models yield very different estimates of the teachers' total contribution to variance as determined by $\sum_{t=0}^{g} \alpha_{gt} \sigma^2_{\theta t}$ and $\sum_{t=0}^{g} \sigma^{T2}_{\theta t}$. For the general model, the teachers' total contribution to variance falls from 138.6 to 78.6 and then to 59.5 or 8.7, 5.6 and 4.5% total variance in scores, as students move from grade 3 to 5. The corresponding estimates for the layered model are 75.7, 167.3 and 200.4 or 4.8, 11.5 and 14.1 percent of the total variance.

The general model suggests that teachers contribute less as students progress through schools, while the layered model suggests the opposite.

Although the models provide very different pictures of the teachers' contribution to variability in scores, they provide reasonably similar estimates of individual teacher's effects. Pooling all three grades of estimated effects, the correlation between the BLUPs from the two models is .69 and is as high as .83 for grade 4 and .91 for grade 5. As suggested by mathematical and simulation-based analysis, BLUPs can be robust to certain model choices such as constraining all $\alpha \equiv 1$. The models also produce similar estimates for the intra-student correlation coefficients. Both models suggest strong intra-student correlation in the residual error terms that is roughly constant at just above 0.8 for grade 3, 4 and 5 test scores.

We also explored the efficiency of estimated teacher effects based on these two models. Of particular interest was the effect that estimating the α coefficients had on the precision of the estimates. The Bayesian framework provides the best format for discussing the precision of estimates while accounting for imprecision in the estimated variance components and other parameters.

Developing a fully Bayesian implementation of these complex models was outside the scope of this study, so we approximated the posterior distribution of the variance components, fixed effects and α's ("**a**" denotes the combined vector of these parameters) by the asymptotic normal distribution of the maximum likelihood estimates. This is the approximate posterior distribution under a non-informative prior distribution (Daniels & Kass, 1998). We then sampled draws from this approximate posterior distribution. Using these draws we calculated the estimated teacher effects $E(\theta \mid Y,$ a) and estimated conditional variance of these effects $V(\theta \mid Y,$ a) for each sampled parameter vector. Across many samples from the approximate posterior, these values lead to Monte Carlo estimates of the posterior mean and variance of θ given the data. In addition, Monte Carlo estimation of quantities such as $P(\theta > 0 \mid Y)$ is straightforward because θ is normally distributed conditional on Y and a.

We found that across teachers the median of the posterior variance was about 17.7 and 12.9 for the layered model for grades 4 and 5 respectively, and 22.4 and 18.5 for the general model by grade. This corresponds to medians for the ratio of posterior to prior variance in teacher effects of .2 and .4 for the layered model for grades 4 and 5 and .3 and .4 for the general model. The posterior variance essentially equals prediction error in the frequentist perspective (Searle et al., 1992). Thus, we have prediction errors that are only between 19% and 39% as large as they would have been had we not used test scores to make our predictions. The estimates are moderately precise and estimating the α's did not greatly degrade the precision of our estimates.

Figure 2 demonstrates how these levels of precision affect inferences about individual teachers. The left panel of Figure 2 is a scatter plot of the posterior probabilities that teacher effects are greater than zero for the layered model plotted against the posterior probabilities for the general model for grade 4 teachers. The right panel

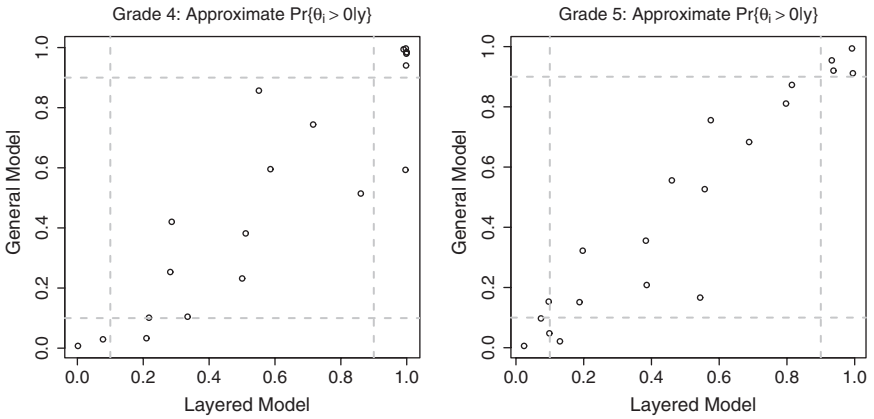

FIGURE 2. *Scatterplots of the posterior probability that the teacher effect is greater than zero for the layered and general models by grade.*

of Figure 2 is the corresponding plot for grade 5 teachers. If imprecision in the estimated teacher effects was very large relative to variability in the true effects, then we would expect to find the estimated probabilities to be clustered around one half. However, this is not what we find. For each grade the estimate probabilities for both models range from near zero to near one with about a quarter to a third of teachers having a high (greater than 0.9) probability of being distinctly different from the mean. Generally, the two models provide very similar results and differences between the posterior probabilities for the two models are often more the results of differences in the estimated teacher effects than in the posterior variances.

As a test of the models on heterogeneous populations, we repeated the study with four of the five charter schools that also serve students from this district.[3] The percent FRL students for the four charter schools, as determined by the cohort in third grade in 1999, is 12, 18, 67 and 79 percent. Percent FRL predicts school and classroom mean scores and gains, even after controlling for individual student FRL status. Almost no students transferred between these schools and thus these charter schools are an example stratification by an omitted covariate.

For these charter schools, Figure 3 plots the estimated teacher effects (BLUPs) against the percentage FRL students in the school. The left panel plots the BLUPs from a model without covariates and shows clearly that estimated teacher effects are strongly related to percentage FRL students. The average BLUPs (marked by **X**) for the two schools serving a low percentage of FRL students are larger than the averages for schools with a high percentage FRL students. Including lunch status as a student-level covariate has almost no effect on the estimates and these estimates are not shown in the figure.

The right panel of Figure 3 plots the BLUPs for a model that includes school fixed effects. In this panel, the estimated BLUPs are essentially uncorrelated with the percentage of students eligible for the lunch program. In addition, by removing vari-

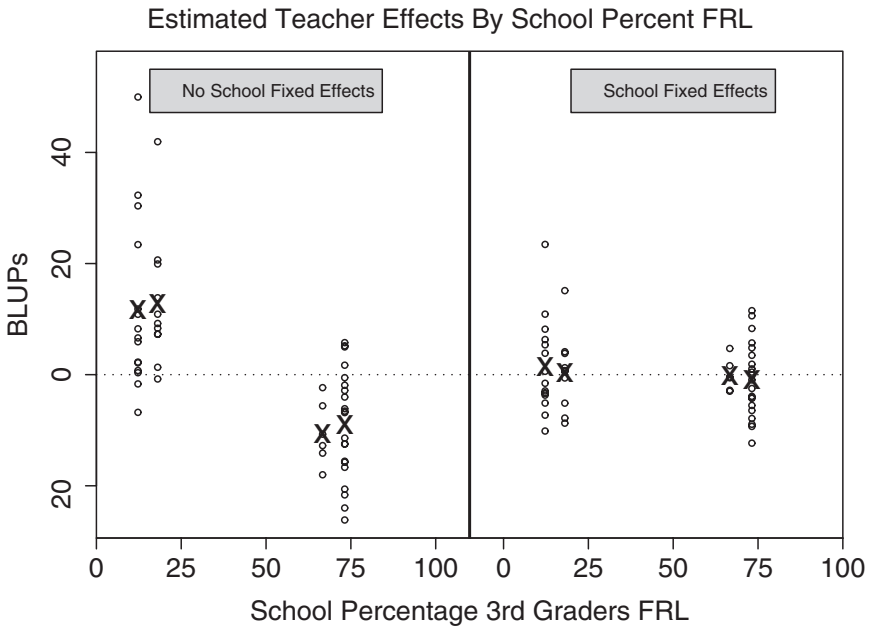

FIGURE 3. *Estimated teacher effects for four charter schools, based on general model with and without school fixed effects.* ***X*** *mark the school mean BLUPs.*

ability related to schools, the fixed effects model estimates smaller teacher variance components resulting in much less dispersion of the BLUPs within schools.

Although including school fixed effects removed the correlation between teacher effects and characteristics of the student population, the model does not disentangle true teacher effects from differences in the student population. In particular, we cannot distinguish what are sometimes called "peer" or "neighborhood" effects where the student population actually affects an individual student's outcomes from heterogeneity in teacher assignments where schools serving lower SES populations attract the least capable teachers. Excluding school effects can bias estimated teacher effects by confounding population effects with true teacher effects. Including school effects can remove legitimate differences in teacher effects among schools. Without additional data, the true source of heterogeneity in scores among schools cannot be determined. Additional covariates might help to disentangle these competing hypotheses but for the most part excessive heterogeneity among the student populations served by schools will result in estimates of teacher effects that are difficult to interpret.

Discussion

The goal of this effort was to explore the issues raised by the application of value-added modeling of student achievement as a means of evaluating teachers

and, by extension, schools. We provide a general model for applying VAM to this purpose and show that all of the currently prominent VAM models of achievement can be seen as restricted cases of this general model. The general model provided a framework for comparing the common models and for evaluating some of the general issues raised by this use of VAM.

Because the computational burden of VAM models has been a concern, it is important to note that fitting the general model proved both feasible and useful for two moderately sized samples of scores from students enrolled in four or five elementary schools from a single large suburban district. The R statistical software run on an off-the-shelf PC workstation was sufficient to maximize the likelihood and provide parameter estimates in a matter of hours.

One difference between our general model and all of the alternative models we reviewed is the assumptions made about the contribution of prior-year teachers to current-year test scores. All of the common models, such as the layered and cross-classified models, assume that these effects persist undiminished over time. In the notation of our general model, they constrain the α parameters to equal one. In contrast, our general model treats the persistence of teacher effects as an unknown and directly estimates the α parameters. The assumption of undiminished teacher effects required by the alternative models is not empirically or theoretically justified and seems on its face not to be entirely plausible. Decaying effects are the norm in much of social science research. In both of the samples noted above, the general model provided a significantly better fit to the data than the layered model, which forces the α parameters to equal one. Therefore the ability to estimate the α parameters directly may be of great use, and further research exploring this issue is needed.

The models also differ in their treatment of the intra-student covariance structure. The cross-classified and the layered models are very similar in this respect, with the layered model allowing more general intra-student covariance structure than the cross-classified model. Estimated teacher effects from both models depend on the adjusted average gain scores for students in their classes, and the adjustments to the gain scores depend on the students' gains from other years. The variability of the total teacher contribution to scores necessarily increases with grade for both these models. However, for many standard assessments, the variance in vertically linked scale scores does not increase with grade. The effects of this possible incompatibility between the model and the data are unknown and warrant further research. An attractive feature of the general model is that because αs can be less than one, the model does not presuppose that variability in the total teacher contribution increases with grade.

One of the major concerns pertaining to all VAM evaluations of teachers is the possibility of bias from the exclusion of covariates, such as student background characteristics. In theory, the omission of covariates that contributes to outcomes can bias parameter estimates when students are stratified by those covariates. This has been a criticism of TVAAS estimates of teacher effects (Ballou, 2002). On the

other hand, individually, both William Sanders and Brian Rowan (personal communications) have argued that adjusting for student-level covariates does not have a great impact on estimated teacher effects. We found that the answer to this argument is complex and depends on the distribution of the omitted covariates and the assignment of students to teachers.

The example in this article is consistent with the comments of Sanders and Rowan that the omission of student-level covariates may not have substantial impact in some cases. And in some cases, modeling the intra-student correlation can mitigate the effects of omitted covariates. However, we found that this is an incomplete answer.

When the student population is stratified into groups that are heterogeneous with respect to the omitted covariates and share few teachers, the omitted covariates will be confounded with teacher effects regardless of the model for intra-student correlation. In the example, adding a student-level FRL indicator to the model had almost no effect on model parameters or estimated teacher effects. However, school average FRL was correlated with school mean scores and gains. Moreover, average FRL predicted scores even after controlling for individual student lunch status. In other words, there were contextual effects for FRL. Controlling for student-level covariates alone is not sufficient to remove the effects of background characteristics in our example and might not be in practical situations. For example, Lee and Bryk (1989) also find contextual effects for student socio-economic status.

Furthermore, controlling for covariates (student or school level) is not as simple as including the covariates in the regression model. When the true teacher effects are correlated with student characteristics, including student covariates in the model can result in systematic errors in estimated teacher effects. The bias arises because the model attributes the true effects of student covariates *and* a portion of the teacher effect to the estimated effect of the covariate and thus wrongfully removes from estimated teacher effect the portion of the true teacher effect that correlates with the covariate. As previously discussed, excluding the covariate from the model has the opposite effect of attributing a portion of the student covariates to the estimated teacher effects.

Although the bias from including student covariates can result whether they are student-level or aggregated to the classroom or school level, controlling for aggregate-level variable and contextual effects poses particular challenges. Some have suggested using the intra-classroom variability of students to provide information for estimating the effects of student-level covariates on test scores and to remove the effects of student-level covariates from estimated teacher effects (Ballou et al., 2003). However, intra-classroom variance cannot be used to estimate contextual effects of aggregate characteristics such as classrooms, schools, peers, and neighborhoods and cannot be used to separate these effects from true differences in the educational effectiveness of teachers and schools. Thus, finding methods for separating contextual effects from teacher effects is likely to be challenging. It may require collecting detailed data on teachers and teacher assignments that generally

are not currently available and may also require making untestable assumptions. However, such methods are important because, heterogeneous populations are likely to be common especially in large school systems and, as our example demonstrates, contextual effects do exist in some school systems.

A key question is the amount of uncertainty in estimates of teacher effects. Numerous observers have asked whether estimates from these models are sufficiently precise to be usable. This paper does not fully address the issue of uncertainty but does shed light on this question. To clarify the implications of this work, however, it is essential to distinguish between model uncertainty and uncertainty stemming from the variability of student scores.

In this article, we used our general model to highlight model uncertainty resulting from assumptions about the persistence of teacher effects, the correlation of scores across grades, and omitted covariates. Relaxing model assumptions by fitting our more general model provides a means of accounting for a portion of model uncertainty that is typically ignored when fitting more restrictive models. The practical impact of model uncertainty will vary from case to case but could be substantial.

Imprecision resulting from variability in student scores will depend on the overall sample size, the sample size per class, and the ratios of teacher and school variance components to the variances component for residual errors. We used our example to study variability in scores. We found that even with only a moderate sized sample, the variability in student scores was sufficiently small so that both the general model and the layered model identified about one fourth to one third of teachers as distinct from the mean.

On the other hand, VAM models will be used to support inferences other than differences from the mean, and the estimates from VAM modeling of achievement will often be too imprecise to support some of the desired inferences. For example, Lockwood, Louis and McCaffrey (2002) found that precise ranking of teachers generally requires the ratio of posterior to prior variance in teacher effects to be very small–no more than 0.1. Thus, in our example, our posterior precision would need to be 2 to 4 times greater to provide meaningful estimates of ranks and accurate identification of teachers in extremes of the distribution.

In general, obtaining sufficiently precise estimates of teacher effects to support ranking is likely to be a challenge. Student test score data tend to be far from ideal, with relatively small classes and substantial numbers of missing values. In addition, models require numerous assumptions that contribute to model uncertainty. Methods to improve precision might include pooling data across years to estimate multi-year average teacher effects. Analysts might also use more restrictive models and ignore the uncertainty in those assumptions–in Bayesian parlance use informative priors to improve posterior variance. More restrictive assumptions about some features of the models, might actually make the models more robust to other assumptions. For example, the general model might be more sensitive to omitted covariates that affect levels but not gain scores than either the layered or CC model.

While this article explores a variety of issues raised by various approaches to VAM estimates of teacher effects, several important issues are not explored. One important issue that is not explored in this paper is the potential impact of missing data. All the models presented in this paper assume noninformative missing data. The sensitivity of estimates of teacher effects to violations of this assumption has not been explored. However, given the large proportion of missing data in many achievement databases and known differences between students with complete and incomplete test data, it is possible that estimates may be highly sensitive to this (or other) assumptions about missing data. This use of VAM also requires assumptions about the consistency of results across the range of scales currently used to measure achievement, as well as other plausible scales that may not be used in VAM efforts undertaken to date. We also noted above that assumptions about the correlations between teacher effects and student characteristics are likely to have an effect on VAM estimates, so further empirical exploration of these relationships and of the impact of assumptions about them could be important.

It is likely that precise estimation of teacher effect will always require assumptions or informative priors. Thus, our work should be interpreted as the first step toward making analysts and consumers of estimated effects aware of the possible impact of model assumptions and toward identifying topics for additional research. By conducting such research we might be able to make well-informed restrictions to models and produce estimates with sufficiently large precision and sufficiently small bias to be useful for making the desired inferences about teachers.

Notes

[1] Brian Rowan (personal communication) reports that when included in the model these interactions had almost no effect on teacher variance components.

[2] If the conditional expectation of y_g given $y_g - 1$ is nonlinear, then the covariate model is not a special case of the general model, which assumes that scores are normally distributed so that conditional expectations is linear. However, the common linear model specification of the covariate adjustment model is a special case of the general model.

[3] We excluded one charter school because of anomalies in the third grade test score data.

Appendix
Intra-Student Correlation and Model Estimates

We focus on a model for the vectors of student gain scores following the discussion above. However, rather than let the ζ_{it}, $t = 1, 2, 3$ have an unspecified covariance structure, we assume $\zeta_{it} = \eta_i + e_{it}$ where the η_i are $N(0, \nu^2)$ and the e_{it} are *iid* random variables with variance σ^2. This model assumes constant correlation among the observations from the same student. For now we will also assume that all the teacher effects are *iid* $N(0, \tau^2)$.

Model A1

$$d_{ij1} = \delta_{i1} + \sum_{l=1}^{N_1} \phi_{ijl}\theta_l + \eta_{ij} + e_{ij}$$

$$d_{ij2} = \delta_{i2} + \sum_{m=1}^{N_2} \phi_{ijm}\theta_m + \eta_{ij} + e_{ij2}$$

$$d_{ij3} = \delta_{i3} + \sum_{m=1}^{N_3} \phi_{ijh}\theta_h + \eta_{ij} + e_{ij3}$$

The main parameter of interest is τ^2 the variance in the teacher effects and will be estimated by the maximum likelihood estimator (MLE). The likelihood is complex and involves matrix inversion that prevents simple analytic evaluation. However, the EM algorithm provides a means of maximizing the likelihood that also provides insight in the contributions of student scores to the estimated teacher effect.

First, we will write Model A1 in vector notation

$$\mathbf{d} = \delta + \mathbf{Z}_1\theta + \mathbf{Z}_2\eta + \mathbf{e}$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are the vectors of teacher and student effects respectively and $\mathbf{Z}_1$ and Z_2 map the specific effect to the correct students and years. The EM algorithm is iterative. At the *mth* step of the algorithm, the current estimates of τ^2 and ν^2 are given by

$$\hat{\tau}^{2[m]} = \frac{t_1}{q_1}$$

$$\hat{\nu}^{2[m]} = \frac{t_2}{q_2}$$

where q_1 equals the total number of teachers across all years and q_2 equals the number of students. We let V denote the variance-covariance matrix for $\mathbf{d}$ as determined by Model A1 and $\mathbf{r}$ equal d less the current estimate of δ. Then t_1 and t_2 are given by:

$$t_1 = \hat{\tau}^{4[m]}\mathbf{r}'\mathbf{V}^{-1}\mathbf{Z}_1\mathbf{Z}_1'\mathbf{V}^{-1}\mathbf{r} + c_1$$

$$t_2 = \hat{\nu}^{4[m]}\mathbf{r}'\mathbf{V}^{-1}\mathbf{Z}_2\mathbf{Z}_2'\mathbf{V}^{-1}\mathbf{r} + c_2$$

where c_1 and c_2 are correction terms that involve only the current estimates of the variance components.

The terms $\mathbf{Z}_1'\mathbf{V}^{-1}\mathrm{r}$ and $\mathbf{Z}_2'\mathbf{V}^{-1}\mathrm{r}$ are the best linear unbiased predictors (BLUPs) of θ and η given the current values for the variance components. Under the assumptions of normally distributed data, the BLUPs are the conditional mean of θ and η given the observed change scores $\mathbf{d}$. Consider a single element of θ, θ_s,

$$
\begin{aligned}
E(\theta_s|\mathbf{d}) &= E\{E(\theta_s|\mathbf{d}, \eta)\} \\
E(\theta_s|\mathbf{d}, \eta) &= \lambda \sum_{is} \phi_{is(t)} \, (d_{it} - \delta_{it} - \eta_i)/n \\
E(\theta_s|\mathbf{d}) &= \lambda \sum_{is} \phi_{is(t)} \, [d_{it} - \delta_{it} - E(\eta_i|\mathbf{d})]/n
\end{aligned}
$$

where $\lambda = \tau^2/(\tau^2 + \sigma^2/n)$ assuming n students per class. At each iteration unknown parameters are replaced by their current estimates.

The BLUP of the teacher effects depends on the conditional mean of the student effects and using a similar approach we find that the conditional mean of η_{ij} is

$$
E(\eta_i|\mathbf{d}) = \omega \sum\nolimits_t \left[d_{it} - \delta_{it} - \sum\nolimits_{is} \phi_{is(t)} E(\theta_s|\mathbf{d}) \right] \Big/ 3
$$

where each student has three scores and $\omega = \nu^2/(\nu^2 + \sigma^2/3)$. Again the student effect depends on the conditional mean of the teacher effect.

Thus, the estimate of the teacher variance component depends on adjusted residuals gains. The gains are adjusted by the fixed time effect and the BLUP of the student effect. The BLUP of the student effect depends on the students' mean gain but is adjusted by the teacher effect estimates from the previous step of the algorithm. In particular, if a student has a, say, large mean residual, but was in classrooms with large mean residuals, the student effect estimate will be much closer to the d_{it} than it would be if there was no classroom heterogeneity in scores. Thus classroom heterogeneity in scores dampens the "shrinkage" in the student effect estimates. Thus, teacher effect variance component estimate is larger than it would be if we used a fixed student effects model. On the other hand, the teacher effect variance component estimate is smaller than it would be if we ignored the correlation among the multiple measures from a single student.

By considering this simple model and the use of the EM algorithm we were able to understand how allowing for correlation among the multiple measures from the same student yields an estimate of the teacher variance component that is a compromise between ignoring the correlation among these measure and a student fixed effect. However, if student gains are related to fixed student characteristics such as socio-economic status or parental education, and these characteristics are heterogeneously distributed across classes, then the random effects model considered here would include the effect of these student characteristics in the teacher effect.

The model we considered was simple but clearly demonstrates how the data are used to estimate the teacher variance component. The contributions of student to teacher effects in the complex layered model will not take on the simple form presented here, but the tradeoff between adjusting for the student and adjusting for his or her teachers will occur. The student effects will differently weight different data from different years but the heuristic notions presented in the appendix apply.

References

Ballou, D. (2002). Sizing up test scores. *Education Next, 2*(2), 10–15.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for students background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37–66.

Burket, G. R. (1984). Response to Hoover. *Educational Measurement: Issues and Practice, 3,*(4) 15–16.

Daniels, M. J., & Kass, R. E. (1998). A note on first-stage approximation in two-stage hierarchical models. *Sankhya, Ser. B, 60,* 19–30.

Diggle, P. J., Liang, K.-Y., & Zeger, S. L. (1996). *Analysis of longitudinal data.* New York: Oxford University Press.

Feldt, L. (1958). A comparison of the precision of three experimental designs employing a concomitant variable. *Psychometrika, 23,* 335–353.

Goldstein, H. (1994). Multilevel cross-classified models. *Sociological Methods & Research, 22*(3), 364–375.

Goldstein, H. (1995). *Multilevel statistical models.* London UK: Arnold.

Hanushek, E. (1992). The trade-off between child quantity and quality. *Journal of Political Economy, 100*(1). 85–117.

Hoover, H. D. (1984a). The most appropriate scores for measuring educational development in the elementary schools: GES. *Educational Measurement: Issues and Practice, 3*(4), 8–14.

Hoover, H. D. (1984b). Rejoinder to Burket. *Educational Measurement: Issues and Practice, 3*(4), 16–18.

Ihaka, R., & Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics, 5,* 299–314.

Lee, V., & Bryk, A. (1989). A multilevel model of the social distribution of high school achievement. *Sociology of Education, 62,* 172–192.

Little, R. S. A., & Rubin, D. B. (1987). *Statistical Analysis with Missing Data.* New York: John Wiley & Sons.

Lockwood, J. R., Louis, T. A., & McCaffrey, D. F. (2002). Uncertainty in rank estimation: Implications for value-added modeling accountability systems. *Journal of Educational and Behavioral Statistics, 27*(3), 255–270.

Meyer, R. (1997). Value-added indicators of school performance: A primer. *Economics of Education Review, 16,* 183–301.

Pederson, E., Faucher, T. A., & Eaton, W. W. (1978). A new perspective on the effects of first grade teachers on children's subsequent adult stat. *Harvard Educational Review, 48*(1), 1–13.

Puma, M. J., Karweit, N., Price, C., Ricciuti, A., Thompson, W., & Vaden-Kiernan, M. (1997). *Prospects: Student Outcomes Final Report* (Technical Report 1997-04-00). Cambridge, MA: Abt Associates, Inc.

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

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

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,* 1525–1567.

Sanders, W., Saxton, A., & Horn, B. (1997). The Tennessee Value-Added Assessment System: A quantitative outcomes-based approach to educational assessment. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluational Measure?* (pp. 137–162). Thousand Oaks, CA: Corwin Press, Inc.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components.* New York: John Wiley & Sons.

Shkolnik, J., Hikawa, H., Suttorp, M., Lockwood, J., Stecher, B., & Bohrnstedt, G. (2002). Appendix D: The relationship between teacher characteristics and student achievement in reduced-size classes: A study of 6 California districts. In G. W. Bohrnstedt, & B. M. Stecher (Eds.), *What we have learned about class size reduction in California Technical Appendix.* Palo Alto, CA: American Institutes for Research.

Snijders, T. A. B., & Bosker, R. J. (1994). Modeled variance in two-level models. *Sociological Methods & Research, 22,* 342–363.

Thum, Y. M. (2002). *Measuring progress towards a goal: Estimating teacher productivity using a multivariate multilevel model for value-added analysis.* A Milken Family Foundation Report. Available at http://www.mff.org/publications/publications.taf.

Webster, W. J., & Mendro, R. L. (1997). The Dallas Value-Added Accountability System. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 81–99). Thousand Oaks, CA: Corwin Press, Inc.

Wertz, C. E., & Linn, R. L. (1970). A general linear model for studying growth. *Psychological Bulletin, 73*(1), 17–22.

Yang, L., & Tsiatis, A. A. (2001). Efficiency study of estimators for a treatment effect in a pretest-posttest trial. *The American Statistician, 55*(4), 314–321.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement, 23,* 299–326.

Authors

DANIEL F. McCAFFREY is Senior Statistician, RAND, 201 North Craig St. Suite 202, Pittsburgh, PA 15213; daniel_mccaffrey@rand.org. His areas of specialization are statistics and education policy.

J. R. LOCKWOOD is Associate Statistician, RAND, 201 North Craig St. Suite 202, Pittsburgh, PA 15213; lockwood@rand.org. His areas of specialization are statistics and education policy.

DANIEL KORETZ is Professor, Harvard Graduate School of Education, Gutman 415 Cambridge, MA 02138; daniel_koretz@harvard.edu. His areas of specialization are educational assessment and measurement.

THOMAS A. LOUIS is Professor, Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe St Baltimore, MD 21205; tlouis@jhsph.edu. His area of specialization is Bayesian statistics.

LAURA HAMILTON is Senior Behavioral Scientist RAND, 201 North Craig St. Suite 202, Pittsburgh, PA 15213; laura_hamilton@rand.org. Her area of specialization is educational assessment.

Manuscript Received March 2003
Revision Received September 2003
Accepted December 2003