*Journal of Educational and Behavioral Statistics*
*June 2007, Vol. 32, No. 2, pp. 125–150*
*DOI: 10.3102/1076998606298039*


# Bayesian Methods for Scalable Multivariate Value-Added Assessment

**J. R. Lockwood**
**Daniel F. McCaffrey**
**Louis T. Mariano**
**Claude Setodji**
*RAND*

*There is increased interest in value-added models relying on longitudinal student-level test score data to isolate teachers' contributions to student achievement. The complex linkage of students to teachers as students progress through grades poses both substantive and computational challenges. This article introduces a multivariate Bayesian formulation of the longitudinal model developed by McCaffrey, Lockwood, Koretz, Louis, and Hamilton (2004) that explicitly parameterizes the long-term effects of past teachers on student outcomes in future years and shows how the Bayesian approach makes estimation feasible even for large data sets. The article presents empirical results using reading and mathematics achievement data from a large urban school district, providing estimates of teacher effect persistence and examining how different assumptions about persistence impact estimated teacher effects. It also examines the impacts of alternative methods of accounting for missing teacher links and of joint versus marginal modeling of reading and mathematics.*

Keywords: *teacher effects; cross-classified multiple-membership data; layered model; Bayesian estimation; Markov Chain Monte Carlo*

## Introduction

Increased testing of students as part of local, state, and federal accountability has resulted in greater availability of longitudinal student achievement data and heightened interest in modeling both student growth and educational inputs to that growth. Value-added assessment (VAA) is one example of such interest in growth modeling (see the Spring 2004 issue of the *Journal of Educational and*

This material is based on work supported by the National Science Foundation under Grant ESI-9986612, the Department of Education Institute of Education Sciences under Grant R305U040005, and the RAND Corporation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of these organizations. We thank Harold C. Doran for providing us the data used in the analyses presented in this article.

*Behavioral Statistics* for a discussion of VAA). To support accountability or educational decision making, VAA models longitudinal test score data with the express interest of estimating the contributions of educators to student learning.

The key feature of longitudinal achievement data for modeling teacher contributions to student achievement is the sequential regrouping of students into different classrooms with different teachers. This results in data where students who are nested under a common teacher for one measurement are not nested together for another measurement. Moreover, scores for students who share a common teacher at one point in time might continue to be positively correlated at subsequent test administrations. The resulting model structures necessary to accommodate these complexities are known as ''multiple-membership'' models (Browne, Goldstein, & Rasbash, 2001; Rasbash & Browne, 2001) because individual scores depend on the effects from multiple ''members'' of the grouping units (e.g., past and current teachers).

Sanders, Saxton, and Horn (1997) developed the ''layered'' model to account for the complex linkage of students to teachers over time and the correlation of future scores for students who shared a past teacher. This model has been used to estimate teacher effects for teachers in Tennessee for many years. Raudenbush and Bryk (2002) proposed an alternative model for longitudinal data with sequential nesting that was used by Rowan and colleagues to estimate teacher effects (Rowan, Correnti, & Miller, 2002). Although the two models differ in important ways (e.g., the layered model accounts for correlation among scores within a student with an unspecified covariance matrix, and Raudenbush and Bryk, 2002, used random growth curves to account for such correlation), the models share a common feature. To account for effects of prior classroom groupings on current and future scores, both models include random teacher effects that are assumed to persist undiminished into all future test administrations. Because of the assumed persistence, we refer to this type of model as the *complete persistence model*.

McCaffrey, Lockwood, Koretz, Louis, and Hamilton (2004) generalized this class of models for longitudinal test scores by explicitly parameterizing and estimating the strength of past teacher effects on current achievement rather than assuming complete persistence. We thus refer to this model as the *variable persistence model*. Using maximum likelihood methods on a small data set, McCaffrey and colleagues estimated the persistence parameters and found them to be positive but substantially smaller than those assumed by the complete persistence model.

As the prospects of VAA for accountability and other high-stakes decisions become more likely, it is crucial to understand to what extent model choices such as how persistence is modeled impact estimated effects. The McCaffrey et al. (2004) finding of small persistence thus motivates the examination of this issue in more realistic data sets; however, the model poses computational challenges that render likelihood methods practically infeasible for all but small data sets. To address this problem, in the Model Development section we present a

Bayesian formulation of the variable persistence model that scales well to the extremely large and complex data sets that challenge alternative approaches to parameter estimation. The formulation includes an extension to jointly modeling outcomes from multiple tested academic subjects (e.g., mathematics and reading) in each year. In the Application and Model Implementation Issues section, we apply these methods to 5 years of student reading and mathematics achievement data from a large urban school district to compare the complete and variable persistence models. This application highlights some challenges that arise when modeling longitudinal test score data and provides a measure of the sensitivity of estimates to alternative possible responses to these challenges, such as how to deal with missing teacher-student links and whether or not to jointly model outcomes from different tested subjects. Finally, the Results and Discussion sections present our findings and offer concluding comments.

## Model Development

### *Specification of Variable Persistence Model for a Single Subject*

McCaffrey et al. (2004) provided a general longitudinal model for $T$ years of annual test score data for a single subject, where throughout this article *subject* refers to an academic content area such as mathematics or reading rather than an individual student. A special case of this model that includes teacher effects but not school effects is the following:

$$Y_{it} = \mu_t + \boldsymbol{\beta}_t' \boldsymbol{x}_{it} + \sum_{t^* \le t} \alpha_{tt^*} \boldsymbol{\phi}_{it^*}' \boldsymbol{\theta}_{t^*} + \varepsilon_{it}. \quad (1)$$

The test scores are denoted $Y_{it}$ for student $i$'s score in year $t$, $t = 1, \ldots, T$. The model includes an overall mean $\mu_t$ for each year, and covariate vectors $\boldsymbol{x}_{it}$ that can include both time invariant and time varying background variables. The model also includes teacher effects $\boldsymbol{\theta}_t$ for each year. For consistency with the literature and for simplicity in presentation we use the term *teacher effects* when describing the random components included at the classroom levels. The effects of interest are not necessarily causal effects or intrinsic characteristics of teachers. Rather, they account for unexplained heterogeneity at the classroom level. Ideally, they provide information about teacher performance, but there might be many sources of this heterogeneity, including omitted student characteristics (McCaffrey et al., 2004; McCaffrey, Lockwood, Koretz, & Hamilton, 2003). Furthermore, the estimated teacher effects are limited in scope to the subject matter aligned to the examination that generated the observed scores and to the appropriateness and validity of the vertical linking procedures employed by the test developer (Reckase, 2004).

The teacher effects are linked to students by the vectors $\boldsymbol{\phi}_{it}$. Elements of these vectors measure the share of the students' instruction provided by each teacher

and take on values in the interval from 0 to 1 (inclusive) with $\Sigma_j \phi_{itj} = 1$. Zeros indicate a student had no instruction from the corresponding teacher, exactly one element equal to 1 indicates the corresponding teacher provided the student with all of his or her instruction, and fractional values indicate a student received instruction from multiple teachers either because of transfer during the year or team teaching. The validity of the chosen fractions would need to be verified for each application, but the model can easily accommodate multiple instructors. For ease of presentation and because our data did not require fractional linkages, for the remainder of the article, we assume that exactly one element of each $\boldsymbol{\phi}_{it}$ equals 1 and the rest are 0.

The value of $\alpha_{tt^*}$ for $t^* = t$ is defined to equal 1 and is used only to make the notation in Equation 1 more compact. The values of $\alpha_{tt^*}$ for $t^* < t$ moderate the covariances of scores in year $t$ among students who shared a teacher at a prior time $t^*$ by modeling the persistence of current teacher effects in subsequent years of testing. A special case of particular interest to the current article is when $\alpha_{tt^*} \equiv 1$ for all $t^* \leq t$, which matches the assumptions made by the complete persistence model. The residual error terms $(\varepsilon_{i1}, \ldots, \varepsilon_{iT})$ are assumed to be multivariate normal with mean vector $\mathbf{0}$ and an unstructured variance-covariance matrix, allowing for different variances at each time point and possibly nonzero and nonconstant correlation of scores from different years (grades), thus making efficient use of the repeated measures at the student level.

### *Multivariate Specification*

The model in Equation 1 is for longitudinal outcomes for a single academic subject such as reading or mathematics. More commonly, students are tested in multiple subjects per year, and it may be of interest to examine teacher effects for these different subjects. Because outcomes for the same student are likely to be positively correlated across subjects, a model for the joint distribution of all outcomes is substantively more appropriate than independent models for different subjects, and exploiting the covariances among all outcomes may lead to reduced bias and improved efficiency in estimates of model parameters (Thum, 1997, 2003).

We generalize the model in Equation 1 to multiple tested subjects per test administration as follows. $\boldsymbol{Y}_i$ denotes the vector of test scores for student $i$. $\boldsymbol{Y}_i$ is of length $ST$, the number of subjects ($S$) times the number of years ($T$), and is organized by subject, and then by test administration (year) within subject. $\boldsymbol{X}_i$ denotes the $(ST \times p)$ design matrix of both time-invariant and time-varying student background variables for the $p$-dimensional vector of regression coefficients

$$\boldsymbol{\mu} = (\mu_{11}, \ldots, \mu_{ST}, \boldsymbol{\beta}'_{11}, \ldots, \boldsymbol{\beta}'_{ST}) \tag{2}$$

where the subscripts on the vectors denote subject and year. That is, the multisubject generalization of the model is constructed as parallel copies of the

single-subject model, with all parameters from the single-subject model being expanded to include separate components for each subject. The teacher effects are expanded along with the parameters of the mean structure so that each teacher has different effects for each subject. We organize the teacher effects by subject and year as $\boldsymbol{\theta}' = (\boldsymbol{\theta}'_{11}, \ldots, \boldsymbol{\theta}'_{1T}, \boldsymbol{\theta}'_{21}, \ldots, \boldsymbol{\theta}'_{2T}, \ldots, \boldsymbol{\theta}'_{S1}, \ldots, \boldsymbol{\theta}'_{ST})$ of length $n_\theta$, where $\boldsymbol{\theta}_{st}$ provides the teacher effects for subject $s$ in year $t$. The matrices $\boldsymbol{\Phi}_i$ generalize to multiple subjects the assignment vectors $\boldsymbol{\phi}_{it}$ by specifying the linkages of students to teachers by subject. Specifically, $\boldsymbol{\Phi}_i$ is $(ST \times n_\theta)$ with only 0 or 1 entries and row sums equal to 1, with the nonzero element in each row corresponding to student $i$'s teacher for a given year and subject. The contribution of teacher effects to the outcomes for student $i$ is then given by $\boldsymbol{A}\boldsymbol{\Phi}_i\boldsymbol{\theta}$, where $\boldsymbol{A}$ is a $(ST \times ST)$ block diagonal matrix consisting of $S$ distinct $(T \times T)$ lower triangular blocks corresponding to subjects. The $(t,t^*)$ element of the block for subject $s$ is $\alpha_{s,tt^*}$ for $t^* \leq t$ and 0 otherwise, where $\alpha_{s,tt'}$ denotes the teacher effect persistence parameters for subject $s$.

The distribution for a single student's score vector, $\boldsymbol{Y}_i$, conditional on the model parameters, teacher effects, and all covariate and linkage information is

$$\boldsymbol{Y}_i|\boldsymbol{\mu}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma} \sim N_{ST}(\boldsymbol{X}_i\boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{\Phi}_i\boldsymbol{\theta}, \boldsymbol{\Sigma}) \tag{3}$$

where $N_{ST}$ denotes the $ST$ dimensional multivariate normal distribution and $\boldsymbol{\Sigma}$ is a $(ST \times ST)$ unstructured positive definite covariance matrix. Outcomes for different individuals are assumed to be conditionally independent given all of these parameters. The teacher effects for each subject $s$ and timepoint $t$ are treated as random effects. We assume that given the variance $\tau^2_{\theta,st}$, the components of $\boldsymbol{\theta}_{st}$ are independent and identically distributed $N(0, \tau^2_{\theta,st})$. The teacher effects are assumed to be independent across subjects and timepoints. More general covariance structures for the random effects (e.g., allowing correlation within teachers across subjects) are possible but are not explored here.

### *Computational Challenges*

Fitting either the single subject or multisubject model to test score data presents considerable computational challenges. Standard mixed-model routines (e.g., those available in R, S-plus, SAS, HLM, and MLWin) are in general not equipped to estimate the persistence parameters. Moreover, even if the persistence parameters are fixed, as in the complete persistence model, the multiple membership structure makes likelihood estimation difficult for realistically sized data sets. Unlike hierarchical models, cross-classified or multiple membership structures prevent reduction of the marginal covariance matrix of the scores into the simple forms that allow for fast computation (Rasbash & Browne, in press). Without these simple forms, likelihood estimation requires inversion of a sparse and generally large matrix. Although specialized likelihood methods for crossed random effects have been proposed (Clayton & Rashbash, 1999; Rasbash & Goldstein, 1994) and

additional methods are in development (Bates & DebRoy, 2004; DebRoy & Bates, 2003; Rasbash & Browne, 2001), to date, only the work of Sanders and colleagues (Ballou, Sanders, & Wright, 2004; Sanders et al., 1997) has been able to scale likelihood estimation of the complete persistence model to large data sets. A scalable implementation of the variable persistence model does not exist.

Recently, Bayesian methods have been proposed as a fruitful alternative approach to the computational challenges presented by crossed random effects (Browne & Draper, 2006; Browne, Draper, Goldstein, & Rasbash, 2002; Browne et al., 2001; Rasbash & Browne, in press; Simonite & Browne, 2003). The utility of Bayesian methods for these problems derives from using Markov Chain Monte Carlo (MCMC) sampling algorithms, particularly successive substitution sampling or "Gibbs" sampling (Carlin & Louis, 2000; Gelman, Carlin, Stern, & Rubin, 1995; Gilks, Richardson, & Spiegelhalter, 1996). The key feature of MCMC algorithms that makes them well suited to handling models with complex relational structures is that sampling the joint posterior distribution of all unknown parameters requires only repeatedly sampling from the conditional posterior distributions of certain parameters given the data and values of all other parameters. Conditioning on random effects reduces the complex covariance matrices to simple, computationally tractable block diagonal forms. Similarly, conditioning facilitates the estimation of the persistence parameters, which can be viewed as unknown regression coefficients of known predictors conditional on the random effects.

*Bayesian Formulation of the Variable Persistence Model*

To capitalize on these advantages, we developed a Bayesian version of the multisubject (or single-subject) persistence model in Equation 3. Given the likelihood, the Bayesian model requires a joint prior distribution for all unknown parameters. The unknown parameters for the model in Equation 3 are $\boldsymbol{\mu}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}$ as well as the variance components $\boldsymbol{\tau}^2 = \{\tau^2_{\theta,st}\}$. We used the following distribution with fixed hyperparameters $\boldsymbol{\mu}_\mu$, $\boldsymbol{V}_\mu$, $\{\delta_{\theta,st}\}$, $\{\nu_{\theta,st}\}$, $\boldsymbol{\mu}_\alpha$, $\boldsymbol{V}_\alpha$, $q_\Sigma$, $\boldsymbol{Q}_\Sigma$:

$$\begin{aligned} p\left(\boldsymbol{\mu}, \boldsymbol{\theta}, \{\tau_{\theta,st}\}, \boldsymbol{\alpha}, \boldsymbol{\Sigma} \mid \boldsymbol{\mu}_\mu, \boldsymbol{V}_\mu, \{\delta_{\theta,st}\}, \{\nu_{\theta,st}\}, \boldsymbol{\mu}_\alpha, \boldsymbol{V}_\alpha, q_\Sigma, \boldsymbol{Q}_\Sigma\right) = \\ N_p(\boldsymbol{\mu}|\boldsymbol{\mu}_\mu, \boldsymbol{V}_\mu) \times \\ \left[\prod_{s=1}^{S}\prod_{t=1}^{T} N_{n_{\theta,st}}(\boldsymbol{\theta}_{st}|\mathbf{0}, \tau^2_{\theta,st}\boldsymbol{I}_{n_{\theta,st}}) U(\tau_{\theta,st}|\nu_{\theta,st}, \delta_{\theta,st})\right] N_{\frac{ST(T-1)}{2}}(\boldsymbol{\alpha}|\boldsymbol{\mu}_\alpha, \boldsymbol{V}_\alpha) \times \\ W(\boldsymbol{\Sigma}^{-1}|q_\Sigma, \boldsymbol{Q}_\Sigma). \end{aligned} \tag{4}$$

Here $N_d$ again denotes a multivariate normal distribution of dimension $d$, $W$ denotes the Wishart distribution, and $U(\cdot\,|\nu, \delta)$ denotes the uniform distribution on the interval $(\nu, \delta)$. Also the counts $n_{\theta,st}$ denote the lengths of the vectors $\boldsymbol{\theta}_{st}$. The joint distribution of the data and parameters is given by the product of Equation 3 across all students and the prior in Equation 4. The posterior distribution of the parameters is proportional to this product. Note that this specification uses

uniform distributions on the standard deviation components rather than gamma distributions on the inverse variance components, the latter leading to closed form full conditional distributions, which are gamma distributions (Gelman et al., 1995). We used the uniform priors because they allow for a somewhat more natural introduction of substantive prior information into the model and also because "noninformative" gamma priors have infinite density at zero, which can lead to undesirable properties of the posterior distribution and associated sampling algorithm.

We implemented this model in WinBUGS (Spiegelhalter, Thomas, & Best, 1999), free software for Bayesian model fitting available at http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml. The WinBUGS implementation was sufficiently fast to estimate the variable persistence model with small data sets and the complete persistence model with small or large data sets. However, fitting the variable persistence model with large data sets or multiple subjects was prohibitively slow. Therefore, we developed an MCMC algorithm (specified in detail in the appendix) for sampling the posterior distribution and programmed it in C, resulting in an efficient and scalable implementation of the multisubject variable persistence model. We use this algorithm for all results reported in this article.

We conducted an extensive simulation study to verify the correctness of the algorithm and its scalability to many teachers and students. We simulated hundreds of data sets from the multisubject persistence model using various configurations of the model parameters, including the persistence parameters, various amounts of missing test scores, and varying numbers of student-level covariates. In all cases the algorithm was able to recover reasonable estimates of the model parameters. Computing times on a standard desktop PC required for the MCMC algorithm to converge sufficiently and to obtain samples for inference were extremely reasonable even for large problems. Depending on the complexity of the simulated data set (i.e., the numbers of students, teachers, years, subjects and regressors, the extent of missing data, and whether or not the $\alpha$ parameters are estimated), the computing times ranged from on the order of several minutes for problems involving a few hundred teachers and several thousand students to a few hours for a huge problem of fitting the complete persistence model for five subjects, 5 years, considerable missing test score data, and following 50,000 students across 50,000 teacher effects (2,000 teachers per subject per year). Fitting the variable persistence model in such a large data set, with five subjects, increases the computing time by about a factor of 6, and we are working on ways to refine the updating of the persistence parameters to reduce this burden. In all cases, RAM requirements are quite low, even for data sets with tens of thousands of students and thousands of teachers.

### *Identifiability of Persistence Parameters*

A technical consideration of the variable persistence model is whether the multiplication of the unknown persistence parameters $\boldsymbol{\alpha}$ and teacher effects $\boldsymbol{\theta}$

makes them unidentified. The issue is further complicated in the Bayesian framework where parameters may be identified only by the prior distribution. Fortunately, assuming that classroom groupings change over time, the data appear to provide sufficient information to uniquely identify $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$. From a theoretical standpoint, the fact that $\alpha_{tt}$ is defined to equal 1 for every $t$ (that is, the teacher effects in year $t$ show up in the model for the year $t$ scores without an unknown coefficient) implies that the signs and scales of the elements of $\boldsymbol{\theta}$ can be estimated without confounding from $\boldsymbol{\alpha}$ and vice versa. Our empirical investigations support this notion. In particular, in the simulations described earlier, the model recovered values of $\boldsymbol{\alpha}$ near 1 when the true values were 1 and values distinctly less than 1 when the true values were less than 1. Moreover, the persistence parameter estimates were not sensitive to prior specification or arbitrary rescaling of the data; this was true in both the simulated data and the actual data used in this study. Finally, we revisited the estimates of persistence obtained via likelihood methods by McCaffrey et al. (2004) and were able to recover those estimates using the Bayesian methods developed here.

## Application and Model Implementation Issues

We fit the multisubject variable persistence and complete persistence models to data from a large urban school district to demonstrate the viability of our Bayesian implementation on actual data, establish empirical evidence on the persistence of teacher effects, and explore the sensitivity of results to alternative assumptions about persistence. The complexities of applying the model to real-world data presented challenges that we needed to resolve to estimate teacher effects under either of the proposed models. Even though they are secondary to the primary goal of understanding the impact of different assumptions about persistence, we provide details on our responses to some of these challenges because they are likely to be faced in any application of VAA. In this section we describe the data, implementation details including specification of prior distributions, and the other practical issues that required consideration.

### *Data Description*

The data contain vertically linked mathematics and reading scale scores on a norm-referenced standardized test administered during the spring of the years 1998 to 2002. For this analysis, we focused on the cohort of students who under normal grade promotion would have been in Grade 1 during the 1997-1998 school year and Grade 5 during the 2001-2002 school year. We thus estimated effects for teachers of Grade 1 during the 1997-1998 school year, Grade 2 during the 1998-1999 school year, Grade 3 during the 1999-2000 school year, Grade 4 during the 2000-2001 school year, and Grade 5 during the 2001-2002 school year. Teacher links to students did not vary by subject; the same teachers were linked to student mathematics and reading scores.

A total of 10,332 students in our data linked to these teachers. However, some of these students had no valid test scores or other problems such as unusual patterns of grades across years that suggested incorrect linking of student records. We deleted records for these students. The final data set includes 9,295 students with 168 unique observation patterns (patterns of missing and observed test scores for both subjects over time), and only about 20% of the students had fully observed scores.

For our analyses we standardized the test scores by subtracting 400 and dividing by 40. We did this to make the variances approximately one and to keep the scores positive with a mean that was consistent with the scale of the variance. Keeping the variance near one made the choice of priors less complex. Keeping the scores positive allowed us to easily code missing values as the only negative values.

*Prior Distributions*

We used the following specifications of the hyperparameters of the prior distributions. $\boldsymbol{\mu}_{\mu}$ was set to the means of the national norming sample for the test, standardized by subtracting 400 and dividing by 40, for consistency with our rescaling of the data. However, the variance $\boldsymbol{V}_{\mu}$ was set to a diagonal matrix of 1,000s to allow the means to be driven by the data. We used independent uniform distributions on the range of $[\nu_{\theta,st}, \delta_{\theta,st}] = [0, 2]$ for the priors for the teacher standard deviation components. Given our rescaling of the data, this range conservatively avoided restricting the teacher variance components. We set $\boldsymbol{\mu}_{\alpha} = \mathbf{1}$ because previous researchers (Ballou et al., 2004; Rowan et al., 2002; Sanders et al., 1997) have assumed this value for their analyses. However, $\boldsymbol{V}_{\alpha}$ was set to a diagonal matrix of 1,000s so these values are determined essentially entirely from the data. $\boldsymbol{Q}_{\Sigma}$ was determined by the inverse of a covariance matrix that assumed a correlation of .7 among student residuals from the same subject, a correlation of .7 between mathematics and reading scores from the same year, and .49 correlation between mathematics and reading scores from different years. These correlations were stylized from analyses we have conducted with numerous student test score databases in our previous research. The standard deviations were set to the values from the national norming sample divided by 40. The degrees of freedom parameter $q_{\Sigma}$ was set to the number of years times the number of subjects plus one (11 in this case). This specification makes the prior minimally informative while maintaining a bounded density over all positive definite covariance matrices.

In addition to these prior specifications, we generated the posterior distribution for the parameters from one model specification using various more informative prior distributions for all parameters as well as overdispersion of the initial teacher effects. The posterior distributions were unaffected as measured by virtually perfect correspondence of the posterior means and standard deviations of all parameters. Finally, as noted previously, gamma priors on the inverse

teacher variance components would lead to gamma full conditional distributions for these parameters. Some of our earlier implementations of the model code used this specification rather than the uniform priors on the teacher standard deviations reported here. The posterior distributions of the model parameters were unaffected by this choice.

*Estimation*

The algorithms previously described and presented in the appendix provide a sample from the posterior distribution for the teacher effects $\boldsymbol{\theta}$ and other model parameters. Any number of features of these distributions such as means, standard deviations, quantiles, and probabilities assigned to different sets might be used to summarize the distributions and compare the parameter estimates across different models. For our comparisons of teacher effects, we focus on two primary summaries of the posterior distributions. We examine the posterior means of the teacher effects $E(\boldsymbol{\theta}|\boldsymbol{y})$, which are the analogs to the best linear unbiased predictors (BLUPs) produced by classical mixed-model estimation procedures. We examine the correlations of the posterior means of the teacher effects across different models, with high correlations indicating that the models are ordering the teacher effects nearly the same. To encompass both central tendencies of the effects and their uncertainties, we also examine $\Pr(\theta > 0|\boldsymbol{y})$, which can be used to identify teacher effects that are "extreme" (either positive or negative) relative to the mean effect of zero. For the comparisons presented in this article, we characterize an effect as extreme and positive if $\Pr(\theta > 0|\boldsymbol{y}) \geq 0.95$ and extreme and negative if $\Pr(\theta > 0|\boldsymbol{y}) \leq 0.05$ (note that given our model for the teacher effects, $\Pr(\theta = 0|\boldsymbol{y}) = 0$). These cutoffs may or may not be appropriate in policy applications, depending on the desired inferences, but are sufficient to quantify the sensitivities that we are exploring. The sensitivity of a teacher effect to a model choice is defined for our purposes as the teacher effect being flagged as extreme under one model and either being flagged as extreme in the opposite direction or not flagged as extreme under another model. These discrepancies should be interpreted with caution as the classifications can be sensitive to small changes in the posterior probability distributions under different models.

For each model that we examined, we used our MCMC algorithm to generate a sequence or chain of parameters sampled from the posterior distribution from that model. We "burned in" each chain for 5,000 iterations and based our inferences on 10,000 post-burn-in iterations. We diagnosed convergence of the chains using the Gelman-Rubin diagnostic (Gelman & Rubin, 1992) implemented in the coda package (Best, Cowles, & Vines, 1995) for the R statistics environment (R Development Core Team, 2005). Using five parallel chains, 5,000 burn-in iterations were clearly sufficient for the ostensible convergence of all the model parameters, including the teacher effects. The posterior means and standard deviations that we report are based on all 10,000 post-burn-in iterations, whereas other inferences regarding teacher effects are based on 1,000 evenly spaced

parameter vectors from the sample of 10,000. Replication of these calculations on different posterior samples obtained from independent chains with dispersed initial values indicated that these posterior sample sizes were sufficient for stability of the inferences.

### *Implementation Issues*

The complexities of the data forced us to address several challenging issues in implementing the model. We discuss four of these: incomplete test score data for a majority of students, missing student-teacher links, nonzero means for estimated teacher effects, and the availability of both reading and mathematics data.

### *Incomplete Test Score Data*

As noted, only 20% of the students had complete testing data for both reading and mathematics over the 5 years for which they were followed. Fortunately, the Bayesian framework provides a particularly simple method for implementing a missing at random (MAR; Little & Rubin, 2002) missing data model known as data augmentation (Schafer, 1997; Tanner & Wong, 1987; van Dyk & Meng, 2001). In essence, missing data are treated as any other unknown parameter, being sampled from their conditional distributions given the observed scores and the values of model parameters (details are provided in the appendix). The inferences about the parameters of interest thus automatically account for the additional uncertainty arising from the missing test score data, provided that the MAR assumption is reasonable.

### *Missing Student-Teacher Links*

Although the Bayesian approach naturally handles missing test score data via data augmentation, the model as specified requires teacher links for all students for all subjects and all years. However, records missing test scores were typically missing teacher information, most likely because the students were not in the school district at the time of testing. How we treated missing teacher links depended on the missingness pattern of the student. Because the accumulation of the teacher effects assumed by the model implies that future teacher effects are not related to past scores, the potential future teacher effects on students who drop out are not required for estimation. We thus assume that all future teacher effects for students who drop out are zero. More care has to be taken with teacher effects that are missing prior to the last observed score for a student. The model for observed test scores depends on these missing links through the persistence of prior teachers, and thus the link must be specified to estimate the model parameters.

We implemented three procedures for dealing with the missing link information. The first and most naive method, which we call *M1* throughout the remainder of the article, assumes that like unobserved future effects for students who drop out, unobserved effects prior to dropout are all zero. The other two methods

assign a teacher effect to each unobserved predropout link that is distinct from the actual teacher effects in the data and occurs at the level of subject within year within student. We refer to these as *pseudo-teacher effects*. The pseudo-teacher effects are treated in the same manner as real teacher effects (*real* referring to actual teachers rather than actual values of the unknown effects), accumulating over time to determine the mean structure for the outcomes. Like the real teacher effects, the pseudo-teacher effects are treated as independent random effects with variance components that vary as a function of year and subject. The two methods that use pseudo-teacher effects differ in their assumptions about these variance components. The simpler method (M2) assumes that the pseudo-teachers for a particular year and subject share the variance component of the observed teachers for that year and subject. This is notionally similar to assuming that the unobserved teacher link is drawn from the population of appropriate teacher effects in the school district, but the pseudo-teacher effect is independent of the real effects. The other method (M3) allows the pseudo-teacher effects to be drawn from their own distributions by year and subject, and the variance components for these distributions are estimated separately from those of the real teachers. This relaxes the assumption that the unobserved effect comes from the same distribution as the real teachers. It reflects the fact that we know nothing about the type of education the students are exposed to when we do not observe them and acknowledges the possibility that these students were in another school district with teachers that differed markedly from the current district. It is important to note that the pseudo-teacher effects are not intrinsically interesting but rather are used only as a tool to obtain more valid estimates of the effects for the real teachers.

### *Nonzero Means for Estimated Teacher Effects*

An additional complication worth noting is that missing teacher links and missing test score data can combine to cause the estimated teacher effects (i.e., the posterior means) to be centered at a small positive value for each grade rather than at zero. This was observed in the actual data, and we reproduced the effect using simulated data as well. The basic intuition behind this phenomenon is that missing test score records, which as noted also tend to be missing teacher links, come from students who tend to score lower. When including the missing teacher effects as part of the total population of teacher effects, which by definition has mean zero by grade, the mean effect for the real teachers is ostensibly positive. This mean shift has no impact on the ordering of real teacher effects but can affect inferences about whether or not effects are positive. We judged that practitioners most often would be interested in knowing whether particular teacher effects were larger than the average teacher effect in the population of teachers being modeled, rather than larger than zero, as the latter question can never really be answered (because the mean structure is always based on the population of students and teachers in the data). Thus, we based our inferences

on teacher effects using the posterior distributions of the centered effects $\theta^* = (\theta - \bar{\theta}_t)$, where $\bar{\theta}_t$ is the average effect of real teachers in year $t$. For example, $E(\theta^*|\boldsymbol{y})$ is the posterior mean teacher effect relative to the average effect for that grade, and $\Pr(\theta^* > 0|\boldsymbol{y})$ is the posterior probability that a teacher effect is better than the average effect for that grade.

*The Availability of Both Reading and Mathematics Data*

A final implementation issue arises from the multivariate structure of the data. We allow each teacher in the data to have a separate effect for mathematics and reading, but these effects can be estimated either by separate univariate longitudinal models (e.g., using Equation 1) fit to each subject or by a joint model (using Equation 3) that makes use of the cross-covariances among the subjects and estimates the effects simultaneously. An interesting and important empirical question is to what extent inferences about the effects may be sensitive to this modeling choice.

Taken in summary, although the primary comparison of interest regards how assumptions about persistence might affect inferences, the empirical issues about how missing teacher links are handled and whether or not the full multivariate structure of the data is exploited are also of interest. We explore these issues by estimating both mathematics and reading effects for all teachers under 12 different settings obtained by a full crossing of the following three factors:

1. complete persistence model ($\boldsymbol{\alpha} \equiv 1$) versus variable persistence model ($\boldsymbol{\alpha}$ estimated),
2. marginal versus joint modeling of the subjects,
3. three different methods for handling missing teacher link information.

## Results

In this section we examine the "secondary" factors regarding missing teacher links and joint versus marginal modeling before considering the issue of persistence. When discussing inferences about the teacher effects, we follow Ballou et al. (2004) in not examining the estimated teacher effects from the first year; in our case, we report only Grades 2 through 5 teacher effects. This restriction is conservative because the layering structures of the models imply that the first year effects are most susceptible to bias by nonrandom assignment of students to teachers.

*Sensitivity to Methods for Missing Teacher Links*

The posterior means $E(\theta^*|\boldsymbol{y})$ for the teacher effects were extremely robust to the assumptions about missing teacher links. For the variable persistence model, the correlations for the three sets of estimates exceeded .99 for every combination of grade, subject, and modeling approach (marginal or joint). Moreover, graphically the estimates align nearly perfectly along the 45 degree line (not

shown), indicating the estimates also share a common variance. Estimates for the other model parameters were also quite robust to the assumptions about missing links, and the models provided very similar estimates of $\Pr(\theta^* > 0|\boldsymbol{y})$. For the complete persistence model, the minimal correlation of estimated teacher effects across missing link methods also was .99. However, there was an additional sensitivity for this model in that missing link method M2 yielded substantially smaller estimates of the teacher variance components than the alternatives for Grade 2 teachers for both mathematics and reading. Fortunately this had a minimal impact on the quantities of interest such as $\Pr(\theta^* > 0|\boldsymbol{y})$.

Given these findings of robustness to the methods for missing teacher links, the remaining results are all based on missing link model M2, which offers a reasonable compromise between the restrictiveness of model M1 and the tenuous information about the variance components of unobserved teacher effects required by model M3. However, the restriction to model M2 did not affect the substantive conclusions of the subsequent results.

*Sensitivity to Joint Versus Marginal Modeling of Subjects*

Jointly modeling the test scores for both subjects versus modeling them marginally with independent replicates of the single-subject model has some effect on the parameter estimates, but the resulting teacher effects are quite similar. The joint model reduces the estimates of the persistence parameters $\alpha$ and the estimates of the teacher variance components $\boldsymbol{\tau}^2 = \{\tau^2_{\theta,st}\}$ for both subjects. The estimated persistence parameters from the marginal models range from about 10% to 80% larger and average 40% larger than those from the joint model. The differences in persistence parameters tend to be largest for the first grade parameters (e.g., $\alpha_{m,21}$, $\alpha_{m,31}$, etc.); otherwise no clear patterns exist in the differences. Similarly, the teacher variance components $\boldsymbol{\tau}^2$ are uniformly larger for the marginal models than for the joint model, with the variance components for the marginal variable persistence model between 10% and 30% larger than those for the joint variable persistence model. The discrepancies in the complete persistence model are somewhat larger, with the ratios of $\boldsymbol{\tau}^2$ from the marginal models to those of the joint model decreasing from about 2.7 for Grade 1 down to 1.1 for Grade 5. We suspect that the differences between these parameters when estimated by marginal rather than joint models results from the joint model exploiting the additional information on the students to more fully separate student effects from teacher effects.

Fortunately, the estimated teacher effects show very little sensitivity to the choice of joint or marginal modeling. The correlation between $\mathrm{E}(\theta^*|\boldsymbol{y})$ under the joint and marginal models exceeds .99 for the variable persistence model and .97 for the complete persistence model for both subjects and all grades. As we would expect, the posterior standard deviations $\mathrm{V}^{1/2}(\theta^*|\boldsymbol{y})$ of the teacher effects are nearly always smaller for the joint model than for the marginal model because pooling information across subjects for students provides more precision

in estimating teacher effects. However, the increase in precision for joint modeling is generally modest, particularly for teachers in later grades. For the variable persistence model, $V^{1/2}(\theta^*|\boldsymbol{y})$ from the marginal model average 7%, 6%, 5%, and 3% larger than those from the joint model for math in Grades 2 through 5, respectively; the analogous percentages for reading are 8%, 4%, 4%, and 2%. The differences in the complete persistence model are more pronounced, with average percentage increases of 20%, 14%, 7%, and 3% for math in Grades 2 through 5 and 16%, 11%, 6%, and 2% for reading in Grades 2 through 5.

The increased precision of the teacher effects under the joint model along with its impacts on the estimates of the other model parameters combine to make the classification of teachers as extreme weakly sensitive to the choice of joint or marginal modeling. Joint models flag somewhat fewer teachers as extreme than do marginal models, and the complete persistence model is slightly more sensitive than the variable persistence model. However, for Grades 2 to 5 these differences are rather small.

*Sensitivity to Assumptions About Teacher Effect Persistence*

The primary goal of our application was to gather additional information about the persistence parameters and sensitivity of results to these parameters. We report our comparison of the complete persistence model ($\boldsymbol{\alpha} \equiv 1$) to the variable persistence model ($\boldsymbol{\alpha}$ estimated from the data) using the joint reading and mathematics model with missing teacher link model M2 (pseudo-teacher effects sharing variance components with real teacher effects). However, as noted, the substantive conclusions hold for all models.

Figure 1 presents posterior means and 95% posterior intervals for the components of $\boldsymbol{\alpha}$ estimated from the variable persistence model. The most important finding is that the $\boldsymbol{\alpha}$ parameters imply long-term persistence of past effects with significant decay in the strength of that persistence over time. All estimates are positive and substantially less than one, with posterior probabilities of being between zero and one approximately equal to one. Thus, the empirical results do not support the assumptions of the complete persistence model. This finding of small, positive estimates of $\boldsymbol{\alpha}$ is consistent with the empirical results presented in McCaffrey et al. (2004) obtained with a different data set.

The disparity between the degree of persistence assumed by the complete persistence model and that estimated by the variable persistence model affects the estimated teacher effects, variance components, and inferences about teachers. Figure 2 plots $E(\theta^*|\boldsymbol{y})$ for reading as estimated by the variable persistence model against the corresponding estimates from the complete persistence model, separately by grade. The corresponding comparisons for the mathematics effects are similar. As shown in the figure, the estimated effects from the alternative models of persistence are moderately to strongly positively correlated, but with notable differences especially in earlier years. Moreover, the data do not lie on the diagonal, except for Grade 5, as the estimates from the variable persistence model

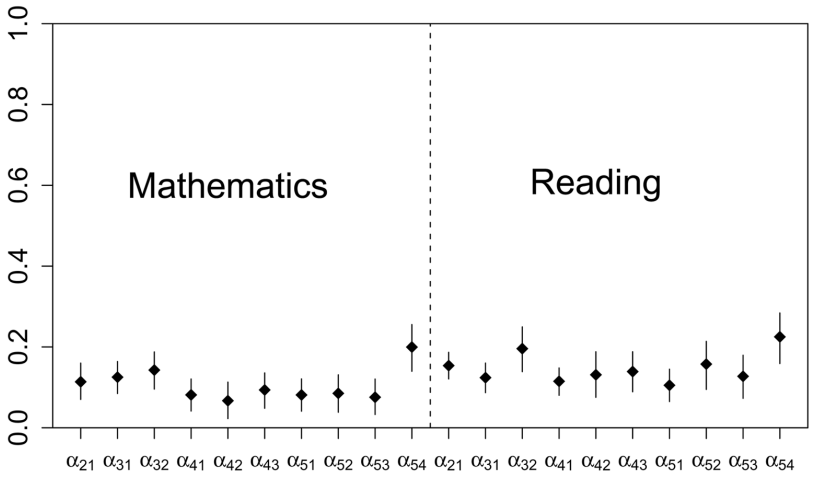


FIGURE 1. *Posterior means and 95% credible intervals for elements of* $\alpha$ *from joint variable persistence model with missing link model M2.*

are substantially more dispersed than the corresponding estimates from the complete persistence model. The difference is most pronounced in Grades 2 and 3, and it is also evident in Figure 3, which presents "caterpillar" plots of third-grade reading teacher effects estimates from the two alternative persistence models. The figures provide a quick, high-level view of the strength of the signal at the teacher level and the rough proportion of teachers whose effects are likely to be different from zero. As clearly shown by the plots, the estimates from the variable persistence model are more variable, and more effects are distinct from the mean.

The increased dispersion in the estimated teacher effects from the variable persistence model results from dramatic differences between the estimated teacher variance components $\boldsymbol{\tau}^2$ from the two models. The posterior means for $\boldsymbol{\tau}^2$ for the variable persistence model are .35, .26, .22, .13, and .17 for mathematics and .41, .16, .18, .13, and .13 for reading, which range from a minimum of 17% to a maximum of 47% of the corresponding posterior means of the residual variance (i.e., the diagonal elements of $\boldsymbol{\Sigma}$) for each year and subject. The analogous estimates for the complete persistence model are .06, .05, .06, .08, and .12 for mathematics and .06, .05, .05, .07, and .10 for reading, which range from 5% to 20% of the corresponding residual variance for each year and subject. The posterior standard deviations of $\boldsymbol{\tau}^2$ are all .04 or less, so the differences are significant. The complete persistence model requires smaller $\boldsymbol{\tau}^2$ in general to maintain consistency with the marginal variance of the data (which is approximately constant over time). The small values of $\boldsymbol{\alpha}$ allow $\boldsymbol{\tau}^2$ in the variable persistence model to be large in general without conflicting with

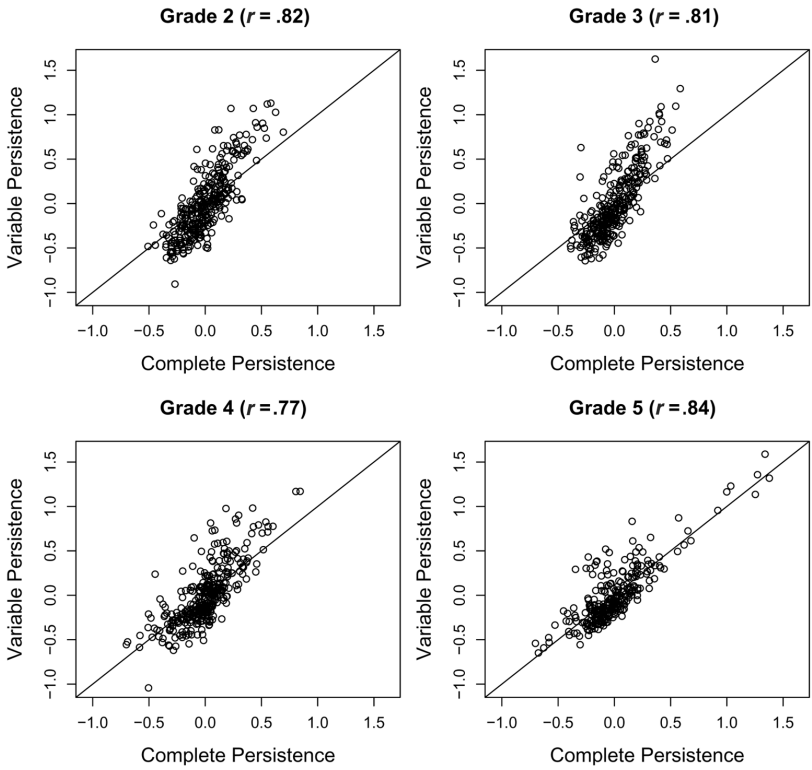


FIGURE 2. *Posterior means of reading teacher effects as estimated by the complete persistence versus the variable persistence models, plotted separately by grade.*
*Note:* The Pearson correlation coefficient is given in the title for each plot, and the diagonal line in each plot represents equality. Effects are based on the joint mathematics and reading models with missing link model M2.

the data. The elements of $\boldsymbol{\Sigma}$ also differ between the two models. In particular, the posterior means of the residual error variances for each subject are substantially larger for the complete persistence model in Grade 1 (0.94 vs. 0.75 for mathematics, 1.15 vs. 0.91 for reading). By Grades 4 and 5, the estimates from the variable persistence model tend to exceed those from the complete model by about 0.04; however, these differences tend to be within the posterior standard deviations, which are about 0.02.

Table 1 summarizes how these differences affect teachers' effects identified as extreme. If the two models perfectly corresponded in their identifications, all counts would fall on the diagonals of the ($3 \times 3$) tables. This is not the case. Generally, the variable persistence model flags substantially more teachers as extreme (both positive and negative), which is consistent with the larger

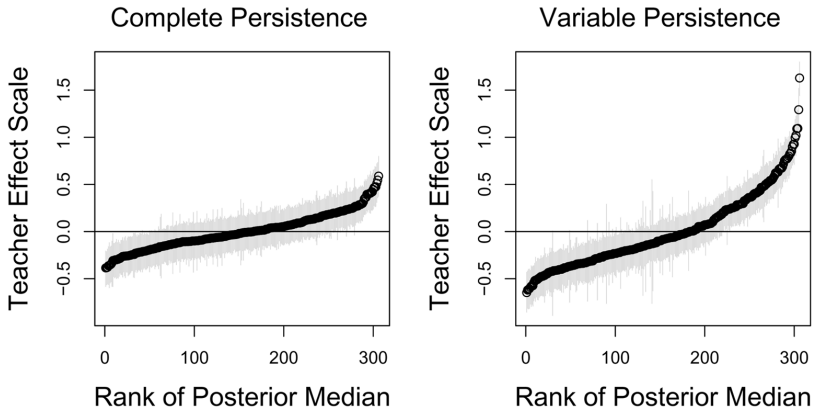


FIGURE 3. *"Caterpillar" plots for Grade 3 reading teacher for the complete (left) and variable (right) persistence models.*
*Note:* Black dots indicate posterior medians, and gray bars delimit 0.05 and 0.95 posterior quantiles. Effects are based on the joint mathematics and reading models with missing link model M2.

estimated teacher variance components. Particularly important is the fact that a small number of teacher effects (given by the upper right-most cells in each table) are flagged as extremely positive by the variable persistence model and extremely negative by the complete persistence model.

The differences in the estimated teacher effects arise because the two models use the data in markedly different ways. As is evident from the model specification, $\alpha_{tt^*} \equiv 1$ implies that gain scores (i.e., $Y_{it} - Y_{i,t-1}$) in a given year are a function of educational inputs only through the current year teacher, and thus gain scores provide most of the information about the teacher effects for the complete persistence model. Alternatively, if $\alpha_{tt^*} \equiv 0$, then past teacher effects do not persist into the future, and current year level scores provide most of the information about current year teacher effects. When $\alpha_{tt^*}$ takes on other values, both level scores and gain scores depend on both current and past teachers, and the model makes more complicated use of the data to estimate effects; however, if $\alpha_{tt^*}$ is very small, then current year levels score are likely to be the primary source of information about teacher effects.

Figure 4 demonstrates how these heuristic notions of the differences between the complete and variable persistence models manifest in our application. The figure is based on reading teacher effects, but the results for math teacher effects are similar. For each teacher, we calculated the average annual reading score for the students linked to that teacher so that each teacher in each grade is associated with five average reading scores summarizing the performance of his or her students, not only in the grade $g$ that the teacher instructs but in each of Grades 1 to 5. We then regressed the estimated teacher effects on these annual

TABLE 1
*Cross-Tabulations of Teachers Flagged as Being Extreme by the Complete and Variable Persistence Models (%)*

| | | Grade 2 ($n = 318$) | | | Grade 3 ($n = 306$) | | | Grade 4 ($n = 321$) | | | Grade 5 ($n = 260$) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | − | 0 | + | − | 0 | + | − | 0 | + | − | 0 | + |
| Math | − | 14 | 2 | 0 | 14 | 3 | 0 | 15 | 3 | 1 | 19 | 4 | 0 |
| | 0 | 20 | 31 | 16 | 19 | 32 | 16 | 13 | 36 | 11 | 13 | 37 | 5 |
| | + | 0 | 3 | 15 | 0 | 2 | 14 | 0 | 7 | 14 | 0 | 5 | 16 |
| Reading | − | 14 | 3 | 0 | 14 | 2 | 1 | 14 | 2 | 1 | 14 | 3 | 3 |
| | 0 | 19 | 37 | 12 | 20 | 36 | 13 | 18 | 40 | 10 | 16 | 41 | 7 |
| | + | 0 | 3 | 12 | 0 | 0 | 14 | 0 | 3 | 13 | 0 | 3 | 12 |

*Note:* For each grade and subject, the (3 × 3) table has rows for the complete persistence model and columns for the variable persistence model. – indicates that the teacher effect is flagged as likely to be negative, 0 indicates that the teacher effect is not detectably different from zero, and + indicates that the teacher effect is flagged as likely to be positive. Perfect correspondence between the models would result in all teachers falling on the diagonal of the table for each subject and grade. Effects are based on the joint mathematics and reading models with missing link model M2.

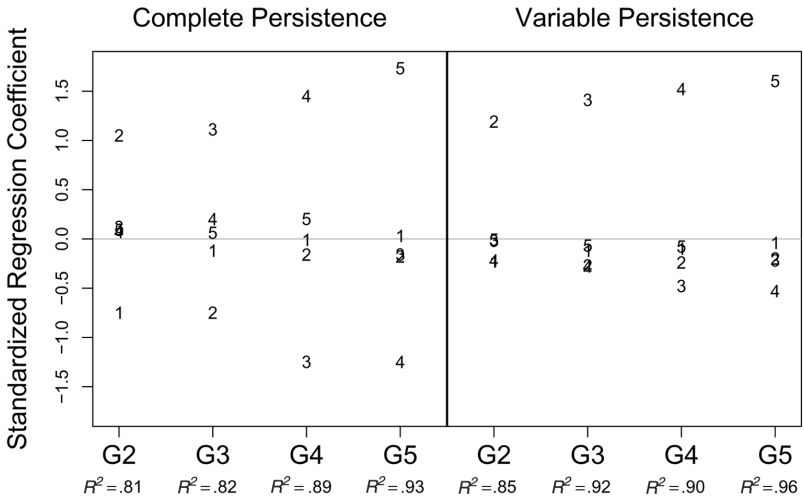


FIGURE 4. *Standardized regression coefficients from regressions of estimated Grade 2 to 5 teacher reading effects on their students' average reading score for every year of testing.*

*Note:* The points are the standardized regression coefficients, with year denoted by the plotting symbol. The $R^2$ for the regressions are printed on the horizontal axis. Effects are based on the joint mathematics and reading models with missing link model M2.

average scores. The figure plots the standardized coefficients by grade and model. The $R^2$ of each regression is presented on the horizontal axis. Although the regression on aggregate scores is substantially less complex than the actual model because it uses only the observed test scores and makes no adjustment for the effects of teachers in other years, the fact that all of the $R^2$ values exceed .8 indicates that the estimated effects are largely identified by simple contrasts. For the complete persistence model, the coefficients for the grade $g$ regression are dominated by a large positive value in grade $g$ and large negative value in grade $g-1$, consistent with the intuition that the complete persistence model identifies teacher effects via average gain scores. On the other hand, the coefficients for the variable persistence model are dominated by the current grade average score. This is consistent with the small positive values of $\alpha_{tt^*}$ estimated from the model.

## Discussion

### *Empirical Findings*

The empirical findings presented in this article using our Bayesian models have several important implications. Our estimates of persistence of teacher effects suggest that scores remain correlated for students who shared a teacher in

the past. This finding is consistent with the hypothesis of persistent teacher effects. However, similar to the results obtained by McCaffrey et al. (2004), the estimates are significantly less than 1 and thus do not support the assumption of complete persistence made by the Sanders et al. (1997) layered model and the Raudenbush and Bryk (2002) cross-classified model. Although the teacher effects from the different models are relatively strongly correlated, the assumptions about persistence have substantive implications for teacher effects being identified as extreme. Further study is necessary to examine which of the models provides "better" estimates of teacher effects as they use the data in notably different ways.

Fortunately, the inferences from the models were largely insensitive to the other implementation choices examined here. Somewhat surprisingly given the degree of missingness in the data, the inferences from the models were almost invariant to the different models for missing teacher links. This robustness may help to alleviate some concerns about the viability of VAA but should be examined in other empirical settings before general conclusions can be made.

Joint versus marginal modeling of the subjects similarly does not appear to exert strong influence on the estimated effects, though the joint models do result in a slight increase in precision for the estimated effects. The fact that the joint model reduces the estimated teacher variance components and persistence is interesting and warrants further consideration; a possible interpretation is that part of what the marginal models are identifying as teacher variance is actually unmeasured student heterogeneity that is more completely removed by the joint models.

Our prior specification for teacher effects assumed that a teacher's effects for reading and mathematics were independent, and alternative priors might provide more precision in the estimated effects. This choice to treat effects as independent prevents potential bias that would result from shrinking the teacher's reading and mathematics effects toward one another when they are disparate, for example, when a teacher effect is significantly larger for one subject than the other. However, the correlation of the posterior means of the effects were .56, .53, .55, and .64 in Grades 2 to 5, respectively, for the complete persistence model and systematically larger at .71, .73, .70, and .76 for the variable persistence model. Clearly the data suggest that teacher effects are correlated and a prior that captures that correlation could allow for sharing information across reading and mathematics, increasing precision of the estimates for both subjects.

There are a few important limitations to the empirical results presented here. First, we do not include any student background variables in our models. As described in McCaffrey et al. (2003), omitted covariates can result in bias, and hence our results might be biased by such variables. We do not adjust for covariates because the current usage of the layered model excludes these variables, and we felt it important to contrast models in a manner consistent with current usage. Also, many data sets that might be used for estimating teacher or school effects might have very limited student measures; for example, the only background

variable included in our data is a race-ethnicity indicator. Future work should include studies that compare the models when covariates are included.

Another limitation of our study is the MAR assumption. There is some evidence (Dunn, Kadane, & Garrow, 2003) that students who miss testing will tend to score lower than others even after controlling for prior test scores. We might find that different models for missing links or the persistence of effects might be more or less sensitive to this assumption. Future work should explore alternative assumptions about missing data.

### *Advantages of the Bayesian Approach*

Although the empirical questions surrounding the persistence of teacher effects ultimately motivated our exploration of Bayesian methods, as shown earlier and as previously noted by Thum (2003), the Bayesian paradigm provides several important benefits for VAA in general. First, as noted, the sequential sampling methods used to make inferences from Bayesian models allow complex cross-classified and multiple membership models for teacher effects to be taken to scale. This advance in capability is practically useful as the demand to perform VAA estimation with large, complicated data sets continues to grow. Second, the Bayesian framework deals naturally with rich latent parametric structures. In a manner similar to the way in which Bayesian estimation methods simplify the treatment of complicated covariance structures, the methods can leverage the sequential sampling approaches to deal successfully with potentially complex functions of multiple latent parameters (e.g., the persistence parameters $\boldsymbol{\alpha}$).

The benefits of the Bayesian framework go beyond its computational advantages. The inferential structure of posterior probability distributions for all parameters, including teacher effects, provides estimates of the uncertainty of the effects of interest that account for uncertainty about all unknown parameters, including variance components. Furthermore, the Bayesian framework facilitates communication of results to policy makers and other consumers of VAA in intuitive terms and straightforwardly allows inferences to be tailored to specific decision analyses that might be required in a given application of VAA. For example, $\Pr(\theta^* > 0|\mathbf{y})$ provides an accessible and simple summary of teacher effects, but inferences about more complex facets of teacher performance, such as Thum's (2003) teacher productivity indices, are similarly easily obtained.

Our future work will include exploring more efficient algorithms for estimating the persistence parameters. In addition, the ease with which the Bayesian framework handles latent parameters provides promise for extending the class of VAA models to include models with richer latent parametric structure. For example, models could include not only latent teacher effects but also latent student effects (akin to the cross-classified model of Raudenbush & Bryk, 2002). The additional latent structure could be used to explore more realistic missing

not at random models for student test scores and to explore models that parameterize teacher effects as functions of latent student effects.

## Appendix: Markov Chain Monte Carlo (MCMC) Algorithm

The full conditional distributions (i.e., the conditional distribution of each parameter or block of parameters given the values of all other parameters, based on the joint posterior distribution) for all of the parameters other than the teacher variance components are known distributions that are easily sampled. (As mentioned in the Model Development section, alternative prior specifications for the teacher variance components that also lead to simple full conditional distributions are possible but are not explored here.) The simplicity of the algorithm derives from the use of residual vectors $\boldsymbol{e}_{\mathrm{i}}$ based on subtracting from $\boldsymbol{y}_i$ all parts of the mean structure that are functions of parameters that are known for a given step of the algorithm. A summary of the algorithm based on initial values of all unknown parameters including the teacher effects is given in the following.

1. Data augmentation: For the observed scores $\boldsymbol{y}_{obs,i}$ for each student, we obtain the observed residuals $\boldsymbol{e}_{obs,i}$ by subtracting the appropriate components of both the fixed effects structure and the random effects structure. Then, under the assumption that scores are MAR (Little & Rubin, 2002), the full conditional distribution of the missing residuals $\boldsymbol{e}_{mis,i}$ is a multivariate normal that depends on only $\boldsymbol{e}_{obs,i}$ and the current value of $\Sigma$. We then add back in the mean structure to obtain imputed values of $\boldsymbol{y}_{mis,i}$. This step results in a fully observed realization of scores $\boldsymbol{y}_i$ and residual $\boldsymbol{e}_{\mathrm{i}}$ for all students.
2. Updating $\Sigma$: First we obtain the residuals $\boldsymbol{e}_i$ for each student. Then the full conditional distribution for $\Sigma^{-1}$ is Wishart depending on the hyperparameters, the number of students, and $\sum_{i=1}^{N} \boldsymbol{e}_i \boldsymbol{e}_i^T$.
3. Updating $\tau_{\theta,st}$: For each subject and year, the full conditional distribution of $\tau_{\theta,st}$ depends only on the current value of $\boldsymbol{\theta}_{st}$, the length of the vector $n_{\theta,st}$ and the hyperparameters $\nu_{\theta,st}$ and $\delta_{\theta,st}$ of the uniform distribution. The full conditional distribution is not available in closed form; we sample a new value for the parameter using the Metropolis-Hastings within Gibbs algorithm (see e.g., Gelman, Carlin, Stern, & Rubin, 1995). This algorithm involves sampling a proposed value from a proposal distribution and accepting that value with a probability that depends on a particular ratio of densities. In implementing this algorithm for the teacher variance components, we update the transformed value

$$\psi_{\theta,st} = \mathrm{logit}\left(\frac{\tau_{\theta,st} - \nu_{\theta,st}}{\delta_{\theta,st} - \tau_{\theta,st}}\right)$$

using the standard change of variables formula to derive its prior density from the prior for $\tau_{\theta,st}$ and using a normal proposal distribution. This parameterization is advantageous because $\psi_{\theta,st}$ is unbounded and thus avoids problems arising from proposing parameters near the boundary of the parameter space.

4. Updating $\boldsymbol{\mu}$: First we obtain the partial residuals $\boldsymbol{e}_i^* = \boldsymbol{y}_i - \boldsymbol{A}\boldsymbol{\Phi}_i\boldsymbol{\theta}$. Then using these residuals, the likelihood and prior are equivalent to a general Bayesian linear model (Lindley & Smith, 1972) for the regression of $\boldsymbol{e}_i^*$ on $\boldsymbol{X}_i$ for all students, with known error covariance $\boldsymbol{V} = (\boldsymbol{I}_N \otimes \boldsymbol{\Sigma})$. The block diagonal structure of $\boldsymbol{V}$ makes the computation of the resulting multivariate normal full conditional distribution for μ particularly simple.
5. Updating $\boldsymbol{\theta}$: We update $\boldsymbol{\theta}$ one element at a time (that is, a single teacher effect for a single subject). We obtain partial residuals $\boldsymbol{e}_i^*$ for every student linked to the teacher of interest based on subtracting the fixed effects structure and the part of the teacher structure that does not depend on the teacher effect being updated. Like the previous step, the resulting likelihood and prior are equivalent to a general Bayesian linear regression model of $\boldsymbol{e}_i^*$ on the single teacher effect θ, where the design matrix consisting of zeros, ones, and the appropriate components of $\boldsymbol{\alpha}$, and where the error covariance matrix is $\boldsymbol{V}$. This results in a univariate normal full conditional distribution for θ.
6. Updating $\boldsymbol{\alpha}$: We update $\boldsymbol{\alpha}$ as a block, across all subjects and time periods. This step is analogous to the previous step, where instead of the $\boldsymbol{\alpha}$ parameters serving as regressors with parameters $\boldsymbol{\theta}$, the $\boldsymbol{\theta}$ serve as regressors with parameters $\boldsymbol{\alpha}$. We obtain partial residuals $\boldsymbol{e}_i^*$ for all students by subtracting the fixed effects structure and the effects of all current year teachers for each score. The resulting likelihood and prior again are equivalent to a general Bayesian linear regression model of $\boldsymbol{e}_i^*$, but now the design matrix consists of zeros and appropriately placed values of $\boldsymbol{\theta}$. The error covariance matrix is again $\boldsymbol{V}$. This results in a multivariate normal full conditional distribution for $\boldsymbol{\alpha}$.

## References


Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, *21*, 37-66.

Bates, D., & DebRoy, S. (2004). Linear mixed models and penalized least squares. *Journal of Multivariate Analysis*, *91*, 1-17.

Best, N., Cowles, K., & Vines, S. (1995). *CODA—Convergence diagnosis and output analysis software for Gibbs sampling output* (Version 0.3). Cambridge, UK: MRC Biostatistics Unit.

Browne, W., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, *1*, 473-514.

Browne, W., Draper, D., Goldstein, H., & Rasbash, J. (2002). Bayesian and likelihood methods for fitting multilevel models with complex level-1 variation. *Computational Statistics and Data Analysis*, *39*, 203-225.

Browne, W., Goldstein, H., & Rasbash, J. (2001). Multiple membership multiple classification (MMMC) models. *Statistical Modelling: An International Journal*, *1*, 103-124.

Carlin, B., & Louis, T. (2000). *Bayes and empirical Bayes methods for data analysis* (2nd ed.). Boca Raton, FL: Chapman and Hall/CRC Press.

Clayton, D., & Rasbash, J. (1999). Estimation in large crossed random-effect models by data augmentation. *Journal of the Royal Statistical Society, Series A: Statistics in Society*, *162*, 425-436.

DebRoy, S., & Bates, D. (2003). *Computational methods for multiple level linear mixed-effects models* (Technical report). Madison: University of Wisconsin–Madison.

Dunn, M., Kadane, J., & Garrow, J. (2003). Comparing the harm done by mobility and class absence: Missing students and missing data. *Journal of Educational and Behavioral Statistics*, *28*, 269-288.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (1995). *Bayesian data analysis*. London: Chapman & Hall.

Gelman, A., & Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457-472.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in practice*. London: Chapman & Hall.

Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear model (with discussion). *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *34*, 1-41.

Little, R., & Rubin, D. (2002). *Statistical analysis with missing data* (2nd ed.). New York: John Wiley.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability, MG-158-EDU*. Santa Monica, CA: RAND.

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

R Development Core Team. (2005). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rasbash, J., & Browne, W. (2001). Modelling non-hierarchical structures. In A. Leyland & H. Goldstein (Eds.), *Multilevel modelling of health statistics* (pp. 93-103). New York: John Wiley.

Rasbash, J., & Browne, W. (in press). Non-hierarchical multilevel models. In J. de Leeuw & I. Kreft (Eds.), *Handbook of quantitative multilevel analysis*. Dordrecht, Netherlands: Kluwer Academic Publishers.

Rasbash, J., & Goldstein, H. (1994). Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model. *Journal of Educational and Behavioral Statistics*, *19*, 337-350.

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

Reckase, M. (2004). The real world is more complicated than we would like. *Journal of Educational and Behavioral Statistics*, *29*, 117-120.

Rowan, B., Correnti, R., & Miller, R. (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.

Schafer, J. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.

Simonite, V., & Browne, W. (2003). Estimation of a large cross-classified multilevel model to study academic achievement in a modular degree course. *Journal of the Royal Statistical Society, Series A: Statistics in Society*, *166*, 119-133.

Spiegelhalter, D., Thomas, A., & Best, N. (1999). *WinBUGS: Bayesian inference using Gibbs sampling* (Technical report). Cambridge, UK: MRC Biostatistics Unit.

Tanner, M., & Wong, W. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, *82*, 528-550.

Thum, Y. (1997). Hierarchical linear models for multivariate outcomes. *Journal of Educational and Behavioral Statistics*, *22*, 77-108.

Thum, Y. (2003). Measuring progress towards a goal: Estimating teacher productivity using a multivariate multilevel model for value-added analysis. *Sociological Methods & Research*, *32*, 153-207.

van Dyk, D., & Meng, X. (2001). The art of data augmentation (with discussion). *Journal of Computational and Graphical Statistics*, *10*, 1-111.

## Authors

J. R. LOCKWOOD is a statistician, RAND Corporation, 4570 Fifth Avenue, Suite 600 Pittsburgh, PA 15213; lockwood@rand.org. His areas of interest include Bayesian methods, longitudinal student achievement modeling, and teacher accountability.

DANIEL F. McCAFFREY is a senior statistician and head of the RAND Statistics Group, RAND Corporation, 4570 Fifth Avenue, Suite 600 Pittsburgh, PA 15213; danielm@rand.org. He has published several papers and reports on value-added modeling of teacher effects.

LOUIS T. MARIANO is a statistician, RAND Corporation, 1200 S. Hayes St., Arlington, VA 22202; Lou_Mariano@rand.org. His research interests include Bayesian hierarchical models, with applications to student assessment and behavioral data.

CLAUDE SETODJI is an associate statistician, RAND Corporation, 1776 Main Street, Santa Monica, CA 90407; setodji@rand.org. His areas of interest are statistics applications in education and health.

Manuscript received September 14, 2004
Accepted July 25, 2005