A Model for Teacher Effects From Longitudinal Data Without Assuming Vertical Scaling
Louis T. Mariano, Daniel F. McCaffrey and J. R. Lockwood
JOURNAL OF EDUCATIONAL AND BEHAVIORAL STATISTICS 2010 35: 253
DOI: 10.3102/1076998609346967

The online version of this article can be found at:
http://jeb.sagepub.com/content/35/3/253

Published on behalf of
American Educational Research Association
and
http://www.sagepublications.com

Additional services and information for Journal of Educational and Behavioral Statistics can be found at:

Email Alerts: http://jebs.aera.net/alerts
Subscriptions: http://jebs.aera.net/subscriptions
Reprints: http://www.aera.net/reprints
Permissions: http://www.aera.net/permissions

>> Version of Record - Aug 13, 2010
What is This?
A Model for Teacher Effects From Longitudinal Data Without Assuming Vertical Scaling

Louis T. Mariano
Daniel F. McCaffrey
J. R. Lockwood
RAND Corporation

There is an increasing interest in using longitudinal measures of student achievement to estimate individual teacher effects. Current multivariate models assume each teacher has a single effect on student outcomes that persists undiminished to all future test administrations (complete persistence [CP]) or can diminish with time but remains perfectly correlated (variable persistence [VP]). However, when state assessments do not use a vertical scale or the evolution of the mix of topics present across a sequence of vertically aligned assessments changes as students advance in school, these assumptions of persistence may not be consistent with the achievement data. We develop the “generalized persistence” (GP) model, a Bayesian multivariate model for estimating teacher effects that accommodates longitudinal data that are not vertically scaled by allowing less than perfect correlation of a teacher’s effects across test administrations. We illustrate the model using mathematics assessment data.

Keywords: teacher effects; value-added models; vertical scaling; Bayesian methods

The No Child Left Behind (NCLB) Act of 2002 required testing of all student in Grades 3–8 and one high school grade in reading and mathematics, by the 2005–2006 school year, and in science by 2008. States and districts are also creating unique student identifiers that allow for linking student achievement test scores across multiple years. Consequently, longitudinal databases of student achievement are increasing rapidly in availability.

These longitudinal achievement data are now commonly used in research on identifying effective teaching practices, measuring the impacts of teacher

This material is based on work supported by the Department of Education Institute of Education Sciences under Grant No. R305U040005 and the RAND Corporation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of these organizations. We thank two anonymous reviewers for helpful comments that improved the clarity of the article, and we thank Harold C. Doran for providing us the data used in the example.
credentialing and training, and evaluating other educational interventions (Bifulco & Ladd, 2004; Gill, Zimmer, Christman, & Blanc, 2007; Goldhaber & Anthony, 2004; Harris & Sass, 2006; Le et al., 2006; Schacter & Thum, 2004; Zimmer et al., 2003). The availability of longitudinal data is also driving interest in using measures of student achievement growth to determine teacher pay and bonuses through the estimation of teacher effects via “value-added” methods (VAMs; Ballou, Sanders, & Wright, 2004; Braun, 2005; Jacob & Lefgren, 2006; Kane, Rockoff, & Staiger, 2006; Lissitz, 2005; McCaffrey, Lockwood, Koretz, & Hamilton, 2003; Sanders, Saxton, & Horn, 1997). There is also intense interest in using value-added modeling to support school decision making (Fitz-Gibbon, 1997) and even suggestions of using the measures to certify teachers or support tenure decisions (Gordon, Kane, & Staiger, 2006).

One of the most challenging aspects of modeling longitudinal achievement data is how to address the persistence of the effects of past educational inputs on future achievement outcomes. In this article, we are concerned primarily with the effects of individual teachers and how best to model the accumulation of those effects across a longitudinal series of student achievement measures. For example, if a teacher improves student reading comprehension by teaching comprehension strategies, then we might expect the strategies to be useful for improving achievement in both current and future years. However, it is less clear how much the effects will persist and how the effects on future achievement will relate to the effects for the current year. The utility of comprehension strategies might diminish over time as students develop other methods for reading comprehension and the teacher’s effect on future scores might decrease and eventually fade to zero.

One approach that has been used when modeling the contributions of past teachers to current and future test scores is to assume that a teacher’s contribution to his or her students’ level of achievement remains constant across test administrations during and after the time the students are assigned to the teacher’s classroom. In essence, a teacher endows a student with a new level of achievement and all future achievements, as measured by the tests, and builds directly on this endowment. For example, if a third-grade teacher’s effect on his or her students in third grade is $+5$ points, the models assume that his or her effect on those students’ scores in fourth, fifth, and higher grades is also $+5$ points. We refer to such models collectively as “complete persistence” (CP) models throughout the article. Examples include cross-classified model of Raudenbush and Bryk (2002), the “layered model” proposed by Sanders et al. (1997) and described in detail by Ballou et al. (2004), and econometric fixed effects models for gain scores (Harris & Sass, 2006).

The assumption of CP is restrictive and might not apply to some data. In particular, if the repeated achievement test scores are not on the same scale of measurement, then it is unlikely that teacher effects would be identical on the different measures for a student. Even if scores are on a single scale of measurement, teacher effects could dissipate or change as students mature and are exposed to additional schooling.
McCaffrey, Lockwood, Koretz, Louis, and Hamilton (2004) and Lockwood, McCaffrey, Mariano, and Setodji (2007) relax the CP assumption by introducing the “variable persistence” (VP) model. This model allows the effect of a teacher on his or her student’s future scores to equal his or her effect from the year she taught the student times a “persistence parameter” that is constant across students and can be estimated from the data. A persistence parameter of less than one indicates that the magnitude of the teacher’s effect is smaller in a future year than when the students were assigned to his or her classroom. If the tests are on a single scale of measure, this could indicate that teacher effects dissipate over time and persistence is incomplete.

However, the VP model still assumes that a teacher’s effect on his or her students’ future test scores is perfectly correlated with his or her effect on their scores when in his or her class—if a teacher’s effect is 2 SD units above the average teacher on the current test, it will be 2 SD units above the mean in all the future years. Again, this assumption may be overly restrictive for some data.

Of particular concern are test scores that are not designed to be on a single developmental scale such as the criteria referenced test used by some states in response to NCLB. For example, the Pennsylvania System of School Assessment, the state’s annual achievement test administered in Grades 3–8 in mathematics and reading, is currently scaled to have a mean of roughly 1,300 and variance of 200 in every grade, without any attempts to link the scores across grades. With such tests, the material tested in later grades may have limited overlap with the material tested in early grades and it may be unrealistic to assume that a teacher’s effects on the material tested in different grades is identical except for scaling constants. In addition, the continued evolution of state assessments implies that at times, developmental scales may be interrupted as an assessment is revalidated for an updated set of curriculum standards.

Even if the repeated measures are linked to single developmental scales, some authors have suggested that the underlying constructs measured by the tests are likely to be multidimensional (Hamilton, McCaffrey, & Koretz, 2006; Reckase, 2004; Schmidt, Houang, & McKnight, 2005). In particular, Schmidt et al. (2005) identify curricular changes across grades and their potential impact on the dimensionality and constructs measured by mathematics tests. Table 1 demonstrates this point. It shows that the number of items pertaining to different facets of mathematical procedures changes substantially across test forms for different grade levels of a standardized test that provides a single developmental scale. For instance, although the testing sequence begins in first grade, two procedural topics, rounding and thinking skills, first appear on the third-grade test, and a third topic, number facts, drops off after the fourth-grade test. The relative weights given to each topic, in terms of the number of items within each, are also shifting from year to year. If a first-grade teacher has somewhat different effects on student learning of these five topics or does not teach rounding or thinking skills, then it would seem likely that the teacher’s effect on a student’s first grade
achievement as measured by this test would not match exactly or be perfectly correlated with his or her effect on the student’s achievement in each of the later years. Lockwood, McCaffrey, Hamilton, (2007) found empirical evidence that teacher effects do differ across different dimensions of mathematics. They estimated teacher effects from two different mathematics subtests, one measuring problem solving and the other measuring procedures. The estimated teacher effects were very weakly correlated, and tests that combined the two different components with different weights would lead to different conclusions about teacher performance.

Hence, there is clearly the potential for the assumptions of both the CP and VP models to be too restrictive for some longitudinal achievement test data. Moreover, overly restrictive assumptions about the persistence of the effects of individual teachers (or other educational inputs) can lead to erroneous inferences. For example, Martineau (2006) shows that when the content mix measured by tests is shifting across grades, models that assume CP can yield severely biased estimate of teacher effects and that bias can also result with the VP model of McCaffrey et al. (J. Martineau, personal communication, October 21, 2004). Lockwood, McCaffrey, Mariano, and Setodji (2007) also showed sensitivity of estimated teacher effects to the specification of the persistence of teacher effects.

In this article, we develop a general multivariate model for estimating teacher effects from longitudinal data, which is flexible enough to accommodate both teacher effect decay and scale changes, including content shift, across tests from different grades. We call the model the “generalized persistence” (GP) model, because its additive structure includes as special cases the CP and VP models referenced above. In essence, each teacher has a vector of effects that includes not only the effect in the grade they teach but also their effects on the future test scores of their current students. These effects are allowed to have an arbitrary covariance structure that is estimated from the data, which relaxes the assumption of perfect correlation of past and future effects made by the existing models.

<table>
<thead>
<tr>
<th>Procedures</th>
<th>Grade 1</th>
<th>Grade 2</th>
<th>Grade 3</th>
<th>Grade 4</th>
<th>Grade 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computation in context</td>
<td>6</td>
<td>8</td>
<td>9</td>
<td>12</td>
<td>16</td>
</tr>
<tr>
<td>Computation with symbols</td>
<td>11</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>10</td>
</tr>
<tr>
<td>Number facts</td>
<td>8</td>
<td>8</td>
<td>6</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Rounding</td>
<td>3</td>
<td>3</td>
<td></td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Thinking skills</td>
<td>9</td>
<td>12</td>
<td>16</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In the GP Model Specification section, we introduce the GP model, show how the existing models are special cases, and contrast how the model uses the data to estimate the teacher effects. In the Estimation section, we present a Bayesian formulation of the model and discuss our choices for prior distributions. We present analyses using actual longitudinal achievement data in the Empirical Example section, considering in particular how the teacher effect estimates obtained from the GP model compare with those from simpler alternatives. Finally, we provide implications and recommendations for areas of future research in the Discussion section.

1. GP Model Specification

We consider longitudinal student achievement outcomes measured over \( T \) time periods. We make a number of simplifying assumptions to improve the clarity of the model description. First, we assume that the data are for a single cohort of students tracked over time. We assume that the periods are consecutive school years and use “years” and “grades” interchangeably when referring to them. We let grade \( g = 1 \) denote the first school year for which student outcomes are available, \( g = 2 \) denote the second school year, and so on. For example, if the first year of testing data is from third grade, then \( g = 1 \) would correspond to third grade, \( g = 2 \) would correspond to fourth grade, and so on. Finally, we assume we are modeling achievement outcomes for a single academic subject such as mathematics. We require that the outcomes be measured on a continuous scale (rather than in categorical performance levels) but the outcomes need not be vertically scaled across grades.

We assume that students are linked at each grade to a teacher responsible for teaching the content on the achievement tests. The goal of the model is to use the longitudinal achievement information combined with student–teacher links to estimate the effects of each teacher. A student’s achievement outcome at each time \( t \) may reflect not only the effect of current teacher but also the effects of past teachers, and we wish to model the effect of each grade \( g \) teacher on their students’ outcomes both in grade \( g \) and in each future year \( t > g \). Because individual scores depend on effects from the student’s membership in multiple group units (e.g., past and current teachers’ classrooms), models like the GP model are generally referred to as “multiple-membership” models (Browne, Goldstein, & Rasbash, 2001; Rasbash & Browne, 2001).

The primary innovation of the GP model is that it allows the current and future effects of a teacher to differ and to be imperfectly correlated with one another. For example, a third-grade teacher has one effect for students when they are in third grade, a different effect for those students when they are in fourth grade, a different effect for those students when they are in fifth grade, and so on. In general, each teacher has a vector of effects for the year they taught a student and for all subsequent years of testing. A grade \( g = 1, \ldots, T \) teacher has \( K_g = T - g + 1 \) effects corresponding to grade \( g \) (the “proximal year”
throughout the article) and all future years up to $T$. For example, if we have scores from Grades 3–8, then a third-grade teacher ($g = 1$) will have six effects, for when his or her students are in Grades 3, 4, 5, 6, 7, and 8, and a Grade 4 teacher ($g = 2$) will have five effects, for when his or her students are in each of Grades 4 through 8, and so on.

We denote the vector of effects for the $j$th teacher ($j = 1, \ldots, J_g$) in Grade $g$ by $\theta_{g[j]}$. The elements of this vector correspond to effects on students’ scores from years $t = g, \ldots, T$. The first element is for the teacher’s effect on students during the proximal year $t = g$ (e.g., the effect of a third-grade teacher on student during third grade). The second element is for the teacher’s effect on students during the year $t = g + 1$ (e.g., the effect of a third-grade teacher on his or her students when they move on to fourth grade) and so on. For this reason, we label the elements of $\theta_{g[j]}$ by $\theta_{g[j]}$. Hence, if we observe data for students in Grades 3–8, the first two elements of the vector of effects for teacher $j$ in Grade 3 are $\theta_{1[j]}$ and $\theta_{1[j]}$, and the first two elements of teacher $j$ in Grade 4 are $\theta_{2[j]}$ and $\theta_{2[j]}$. We use this somewhat nonstandard notation so that labeling of the elements corresponds to the student’s grade, regardless of which grade the teacher actually taught. Table 2 gives an example of this notation for teacher effects experienced by an individual student progressing from Grades 3–8.

Our model assumes that a student’s year $t$ score depends on an overall year $t$ mean for all students, plus a cumulative sum of the current year and past year teachers’ effects, plus a random residual error term for the student in this year of testing. Let $y_{it}$ be the achievement score of student $i$ in year $t$; our model for this score is

\[
y_{it} = \mu_t + \sum_{j=1}^{J_g} \theta_{g[j]} y_{jt} + \varepsilon_{it}
\]
\[
y_{it} = \mu_t + \left( \sum_{g=1}^{t} \sum_{j=1}^{J_g} \phi_{igj} \theta_{g[j]} \right) + \varepsilon_{it}.
\]

where \(\mu_t\) is the overall mean for the year, \(\phi_{igj}\) equals 1 if student was taught by teacher \(j\) in year \(g\) and 0 otherwise, so that the products \(\phi_{igj} \theta_{g[j]}\) provide the teacher effects for the current and prior grades teachers, and \(\varepsilon_{it}\) is the residual error term. For simplicity, we assume that each student had only one teacher each year, but \(\phi_{igj}\) could be modified to allow for multiple teachers and teachers with part-year contributions.

The residual error terms \(\varepsilon_{i}' = (\varepsilon_{i1}, \ldots, \varepsilon_{iT})\) are assumed to be normally distributed random variables, independent across students, with a mean of 0 and an unstructured covariance matrix \(\Sigma\):

\[
\varepsilon_{i} \sim MVN(0, \Sigma).
\]

For each \(g\), we model the grade \(g\) teacher effects with a \(K_g\)-dimensional multivariate normal distribution with mean vector \(0\) and unstructured covariance matrix \(\Gamma_g\):

\[
\theta'_{g[j]} \sim MVN(0, \Gamma_g).
\]

(For \(g = T\), this reduces to univariate normal because \(K_g = 1\).) We assume that the vectors of teachers’ effects are independent across both teachers in the same grade and teachers in different grades and are also independent of the student residual errors.

Equations 1, 2, and 3 comprise the basic GP model for estimating teacher effects from longitudinal data on a single academic subject. Following McCaffrey et al. (2004), the model may be extended to account for time invariant and time varying student background variables \(x_{it}\) and school effects \(\eta_{g[k]}\) (indexed analogously to \(\theta_{g[j]}\)) as

\[
y_{it} = \mu_t + \mathbf{\beta}x_{it} + \left( \sum_{g=1}^{t} \sum_{j=1}^{J_g} \phi_{igj} \theta_{g[j]} \right) + \left( \sum_{g=1}^{t} \sum_{j=1}^{K_g} \lambda_{igk} \eta_{g[k]} \right) + \varepsilon_{it}.
\]

The vector \(\mathbf{\beta}\) parameterizes the effects of each student background variable. Students are linked to schools by the indicators \(\lambda_{igk}\), which functions in the same manner as \(\phi_{igj}\). We concentrate on the more parsimonious version of the model described in Equation 1 that focuses on teacher effects only.

1.1. Special Cases: Complete, Variable, and Zero Persistence

We chose the name “generalized persistence” model, because it includes as special cases other models for longitudinal achievement data that make more restrictive assumptions about the persistence of teacher effects. Here, we discuss
three special cases: the “zero persistence” (ZP) model, where a teacher’s effect is assumed to not carry over at all into future years \( (\theta_{g|t|} = 0, \ t > g) \); the CP model, where a teacher’s effect is assumed to carry forward undiminished and unchanged to all future years \( (\theta_{g|t|} = \theta_{g|g|}, \ t > g) \); and the VP model, where a teacher’s future effects are simple rescalings of the proximal year effect, with separate rescaling factor for each year, but future effects are otherwise unchanged \( (\theta_{g|t|} = a_g \theta_{g|g|}, \ t > g) \).

It is also useful to characterize the various models in terms of their assumptions about the covariance matrices \( \{\Gamma_g\} \). We decompose \( \Gamma_g = S_g^{1/2} C_g S_g^{1/2} \), where \( S_g \) is a nonnegative diagonal matrix of the variances of the grade \( g \) teacher effects in each outcome year \( t \geq g \) and \( C_g \) is the nonnegative definite correlation matrix of those effects. The GP model is “general” in the sense that it places no constraints on either \( S_g \) or \( C_g \)—the variances of the grade \( g \) teacher effects are allowed to differ for each \( t \geq g \) and the correlation structure of these effects is arbitrary. Consequently, \( \Gamma_g \) has \( (K_g)(K_g + 1)/2 \) free parameters—the \( K_g \) variances in \( S_g \) and the \( (K_g)(K_g - 1)/2 \) correlations in \( C_g \).

The other models are special cases because their \( \{\Gamma_g\} \) have fewer free parameters. The ZP model restricts \( \Gamma_g \) to one free parameter, the variance \( \tau_g^2 \) of the proximal year teacher effects. With no carry over of effects to future years, all other elements of \( \Gamma_g \) are 0. The CP model also restricts \( \Gamma_g \) to one free parameter, the variance \( \tau_g^2 \) of the proximal year teacher effects, but because those effects are assumed to persist indefinitely, \( S_g = \tau_g^2 I \) and \( C_g = J \), a matrix of all 1s. Finally, in the VP model each \( \Gamma_g \) has \( K_g \) free parameters, the diagonal elements of \( S_g \), and \( C_g \) is again constrained to equal \( J \) because the future year effects as rescaled proximal year effects are perfectly correlated with the proximal year effects themselves.

One conceptually appealing property of the VP model is that, assuming a unidimensional construct measured by each test, the persistence of teacher effects on achievement as measured by the tests is explicitly parameterized by \( a_g \). Once we allow for the possibility of less than perfect correlation of teacher effects to accommodate more complex relationships among the constructs measured by the various tests, there is no direct analog in the more flexible GP model to that of the VP model’s persistence parameters. This is because in restricting the correlation to \( J \), the VP model makes a strict assumption about how the teacher effects may change that forces any decrease in persistence to be fully absorbed into \( S \). Once that assumption is lifted in the GP model, both the variance and correlation elements of \( \Gamma_g \) may be affected by the nature of the true persistence and there is no one GP parameter that fully encapsulates persistence.

### 1.2. How the GP Model Uses Data to Estimate Teacher Effects

Because the GP model allows teacher effects to persist into the future, the complexity of the model obscures how the data contribute to estimated effects.
In this section, we explore how test scores from each year (past, proximal, and future) contribute to the estimation of teacher effects under different values for the covariance matrices of teacher effects \{\Gamma_g\} including values that correspond to the special cases of the CP, VP, and ZP models.

Estimates of teacher effects using the GP model likelihood (via maximum likelihood and mixed model estimators or using the Bayesian methods we propose in the next section) have the form \( \theta = Qr \), where \( Q \) is a function of \( \Sigma \), \( \Gamma_g \), and the \( \phi_{ig} \) in Equation 1 (Searle, Casella, & McCulloch, 1992), and \( r \) is the vector of residuals of the test scores \( y \) away from any fixed effects (e.g., annual means) in the model. Each row of \( Q \) provides the weights given to the elements of \( r \) when producing the estimate for a particular element of \( \theta \) and \( Q \) fully characterizes the use of data by the GP model in estimating teacher effects. We study how different values for the covariance matrices, \( \{\Gamma_g\} \), lead to changes in \( Q \) and the use of data in estimation of \( \theta \).

We conduct an empirical exploration of \( Q \) using a simple simulated data scenario, where 125 students are followed for 5 years, and each year the students are randomly regrouped into five classes of size 25. The covariance matrix for each student’s residual errors, \( \Sigma \) is assumed to be a five by five matrix with ones on the main diagonal and 0.7 for all off-diagonal elements (i.e., is compound symmetric). Variations in \( Q \) depend only on our assumptions about the \( \{\Gamma_g\} \), and we consider seven different cases for these matrices. For each case, every \( \Gamma_g = \tau^2S^{1/2}(\alpha)C(\rho)S^{1/2}(\alpha) \), where \( C(\rho) \) denotes a compound symmetric correlation matrix of conforming dimension with correlation \( \rho \) in every off-diagonal cell, and \( S(\alpha) \) denotes a conforming diagonal matrix with elements \((1, \alpha, \alpha^2, \alpha^3, \ldots)\). The parameter \( \tau^2 \) equals the marginal variance of teacher effects in the proximal year, and we set it to 0.25, which implies that the variability among proximal year teacher effects is one fourth as large as the corresponding residual variance for students.

The different cases we consider vary only on the values of \( \alpha \) and \( \rho \) and are as follows: “CP” (\( \rho = 1, \alpha = 1 \)), “GP1” (\( \rho = 0.7, \alpha = 1 \)), “ZC1” (\( \rho = 0, \alpha = 1 \)), “VP” (\( \rho = 1, \alpha = 0.5 \)), “GP2” (\( \rho = 0.7, \alpha = 0.5 \)), “ZC2” (\( \rho = 0, \alpha = 0.5 \)), and “ZP” (\( \rho = 0, \alpha = 0 \)). As the notation indicates, \( \rho = 1, \alpha = 1 \) corresponds to the CP model; \( \rho = 1, \alpha = 0.5 \) corresponds to the VP model with persistence parameters defined by \( S(0.5) \); and \( \alpha = 0 \) corresponds to the ZP model. Opposite the CP and VP models, the ZC1 and ZC2 models are GP models, where a teachers’ effects in the proximal and each future year are independent. In the GP1 and GP2 models, the correlation of teacher effects is less than 1 but greater than 0. The CP, GP1, and ZC1 cases have the property that for each \( \{\Gamma_g\} \), all of the diagonal elements are equal, so that the variabilities of teacher effects in the proximal and all future years are the same. The VP, GP2, and ZC2 cases have decreasing variance of the future year effects relative to the proximal year, with ZP being the extreme case of zero variance of the future year effects.
Figure 1 summarizes the comparisons of the different models. Because the simulation was completely balanced and students were randomly assigned to teachers, any variation in the rows of $Q$ across the five teachers within a grade is random, and so it is sufficient to consider only a single teacher of each grade to represent their grade. This makes a total of 15 different fundamental types of teacher effects (the five effects $\theta_{1[j]}$, ..., $\theta_{1[j]}$ of a Grade 1 teacher, the four effects $\theta_{2[j]}, ..., \theta_{2[j]}$ of a Grade 2 teacher, and so on, to the single effect $\theta_{5[j]}$ of a Grade 5 teacher). Of these 15 types of effects, Figure 1 displays four ($\theta_{2[j]}, \theta_{2[j]}, \theta_{3[j]}, \theta_{3[j]}$) as canonical; the patterns exhibited by the other effects generalize naturally. Each frame of Figure 1 corresponds to a different effect, with a summary of $Q$ for each of our seven cases of the GP model. For each teacher effect and case, we plot five different points that correspond to the average weights given by $Q$ to the residuals from the five different years for the students who are linked to that teacher. For example, in the upper left frame corresponding to $\theta_{2[j]}$, for each case, “1” equals the average weight given to the
Grade 1 (prior year) residuals for the students who are linked to this second-grade teacher in Grade 2, “2” equals the average weight given to the Grade 2 (proximal year) residuals for these same students, “3” equals the average weight given to the Grade 3 (future year) residuals for these students, and similarly for “4” and “5.” In essence, the plots indicate for a given teacher how the scores from every year for the students linked to that teacher are contrasted to produce the estimate of that teacher’s effect.

As shown in the figure, the scores for years prior to the proximal year receive negative weights in all cases (except for future year effects in the ZP model, where all scores receive a weight of 0). Negative weighting of prior scores is consistent with value-added modeling’s goal of adjusting for students’ past test scores when estimating current teacher effects. For instance, if simple mean gains were used to estimate teacher effects, then the scores from the previous year would be given weight equal to the negative of the weight given to the current scores. The additional structure of the GP model makes these weights more complex, but they retain the same essential logic.

One notable feature about the weighting of data by the GP model is the effect of assuming perfect correlation among a teacher’s effects across time (CP and VP models). For estimates of future year effects, perfect correlation results in the proximal year having the largest positive weight, whereas correlation less than 1 results in the future year having the largest weight (e.g., the difference in weights given to for Year 5 scores for estimating $\theta_{2|5}$ or $\theta_{3|5}$ using VP and CP models compared with the other models). When the correlation is 1, the proximal year contains all the same information as the future years and the data from the proximal year are more informative because they involve fewer intervening effects. For $\theta_{2|5}$ under VP, the Year 5 data actually receive negative weight because the small $\alpha$ value implies that the Year 5 scores provide more information about the students’ general performance levels than the teacher’s effect, and hence, it is more efficient to treat the Year 5 scores like prior scores than as positive information about the Grade 2 teacher. When the correlation is less than one, then future scores provide the only direct estimate of the future effect and they receive large positive weight. Simulations not shown here revealed a smooth but rapid transition from giving the largest weight to the proximal year to giving the largest weight to the future year. Even with a correlation as large as .95, the future year received greater weight than the proximal year. A comparison of GP1 with ZC1 or GP2 with ZC2 shows that further decreases in the correlation have little effect on the weighting of the data for estimating teacher effects.

Another notable feature of the weighting is that allowing the variance to decline compresses the weights for estimating future year estimates (so as to match the scale), and it creates greater relative spread in the weights for years that are not of interest (e.g., Years 3 to 5 in $\theta_{2|2}$ or Years 2 to 4 for $\theta_{2|5}$). This disparity in the weights occurs because decreasing the variance implies that the
intervening years contain relatively more student information than they do when the variance is constant and this shifts how the data are used for sorting out teacher and student inputs.

It is clear that, relative to the GP model, the models that have been used in the past (CP and VP) make very specialized use of the data and give more weight to the proximal year data than future year data in future year estimates. It is also clear that allowing the variance of the teacher effects to decline in future years allows for differential weighting of future year data, whereas forcing the variance to be constant forces all future years to have roughly equal weight in the proximal year estimates.

2. Estimation

As discussed by Lockwood, McCaffrey, Mariano, and Setodji (2007), the Bayesian framework (Carlin & Louis, 2000; Gelman, Carlin, Stern, & Rubin, 1995; Gilks, Richardson, & Spiegelhalter, 1996) has a number of advantages for estimating teacher effects from complex longitudinal models, including computational efficiency in the presence of the multiple-membership data structure and a natural mechanism for making inferences about individual teacher effects and other parameters of interest. We thus implement the GP model in the Bayesian framework. In this section, we discuss our choice of prior distributions and implementation of the model in the Bayesian modeling software WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000).

A complete Bayesian specification of the basic GP model (Equations 1, 2, and 3) requires prior distributions for the unknown parameters, $\mu_1, \ldots, \mu_T, \Sigma$, and the $\Gamma_g$. To allow the data to drive the parameter estimates, we use independent, minimally informative, natural semiconjugate (Gelman et al., 1995) priors of very low precision normal distributions for each $\mu_t$ and a Wishart distribution for $\Sigma^{-1}$ with $T + 1$ degrees of freedom.

For each $g$, we assume $\Gamma_g$ is distributed Wishart with $g + 1$ degrees of freedom and a scale matrix where $\text{cor}(\theta_{g|j|t}, \theta_{g|j'|t'}) = 0.8$, for $t \neq t'$ and $\text{var}(\theta_{g|j|t}) = 0.25 \times 0.7^{(t-g)}$. The “0.25” term in the variance implies that the proximal year teachers account for about 20% of the total variance in student scores. We did not assume that $\Gamma_g^{-1}$ is distributed Wishart, even though this is the semiconjugate prior, because we found that assuming $\Gamma_g$ is distributed Wishart better matched our desire for the prior distribution to be minimally informative and to place nontrivial mass on matrices for which $\text{cor}(\theta_{g|j|t}, \theta_{g|j'|t'})$ is large for $t \neq t'$ and for which $\text{var}(\theta_{g|j|t})$ decreases with $t$. We needed to be more explicit when choosing this prior because previous research (Barnard, McCulloch, & Meng, 2000) has shown that it is difficult to select prior distributions that are minimally informative for all parameters of covariance matrix. Hence, we...
wanted our prior to be consistent with previous research that suggests that variation in the future year effects of teachers is notably smaller than the variance in the proximal year (Lockwood, McCaffrey, Mariano, & Setodji, 2007) and that has implicitly assumed \( \text{cor}(\theta_{g|t}, \theta_{g|t'}) = 1 \) through the use of the GP and VP models.

Through simulation studies, we determined that modeling \( \Gamma_g \) was satisfactory at meeting these objectives and could be implemented in WinBUGS. We simulated data sets of 2,500 students followed for 5 years, where each year students were randomly regrouped into classrooms of size 25 (so that there are 100 teachers per year). Data sets were simulated under different instances of the GP model by varying the values of \( \rho \) and \( \alpha \) discussed in a previous section. For each data set, we fit the GP model in WinBUGS using a minimally informative Wishart prior for \( \Gamma_g \), and we used samples from the posterior distribution to estimate the posterior means of all model parameters including the elements of each \( \Gamma_g \) and functions of those elements, most notably the correlations. We then compared the posterior means to the true values used in generating the data. In all cases, the posterior means were consistent with the true values, even in cases where \( \Gamma_g \) had high correlations (\( \rho = 0.85 \)) and declining variance of future year effects relative to the proximal year (\( \alpha = 0.6 \)). Overall, these simulations suggest that even though the scale matrix of our Wishart prior for \( \Gamma_g \) was chosen to match prior research, the distribution was sufficiently diffuse to support either high or low correlations, regardless of the scaling of the variance components.

However, our simulations corroborated previous findings that a minimally informative Wishart prior for \( \Gamma_g^{-1} \) is actually quite informative for correlation parameters when the diagonal elements of the covariance matrix differ considerably in size. Using this prior with our simulated data resulted in posterior means that in many cases were systematically different than the true values. In particular, when \( \rho = 0.85 \) and \( \alpha = 0.6 \), the Wishart prior for \( \Gamma_g^{-1} \) led to posterior means of the correlations that were in the range of 0.5 to 0.7, much smaller than the true value of 0.85, and for most values of \( g \) and \( t \), the upper bounds of the 95% posterior credible intervals for the correlations did not include 0.85. Auxiliary investigations indicated that particularly when \( \alpha < 1 \), it was virtually impossible to specify a Wishart prior for \( \Gamma_g^{-1} \) that did not substantially down-weight high correlations unless we were willing to be extremely informative that correlations were indeed high.

The main challenge with specifying a Wishart prior distribution for \( \Gamma_g \) was computational. As noted, we implemented the model in WinBUGS (Lunn et al., 2000), a freely available program that has a language for specifying Bayesian models and implements those models using Markov chain Monte Carlo (MCMC) methods (Carlin & Louis, 2000; Gelman et al., 1995; Gilks et al., 1996) to obtain samples from the model posterior distributions. In the most current version of WinBUGS (1.4), the built-in function for the Wishart prior
distribution can be used only for the inverse covariance matrix of a multivariate normal distribution. Consequently, when specifying our prior distribution for $\Gamma_g$, we used the result that a Wishart distributed random matrix with degrees of freedom equal to $g + 1$ and a specified scale matrix can be generated as the scaled square of a random $g$-dimensional lower triangular matrix, where subdiagonal elements are standard normal variables and the diagonal elements are chi-square variables with degrees of freedom decreasing from $g + 1$ to 1 (Smith & Hocking, 1972). Using this approach had the side effect of reducing the computational efficiency of the algorithm in general. In some cases, inefficient starting values of $\theta$ would cause the program to run slowly during the initial MCMC iterations; however, allowing WinBUGS to randomly generate the $\theta$ starting values avoids this issue. The substantive advantages of this prior distribution outweighed the additional computational burden in our application. Example code for fitting the model to our data discussed in the Empirical Example section is available electronically from the JEBS Web site (http://jeb.sagepub.com/ supplemental).

3. Empirical Example

In this section, we present an application of the GP model to actual student achievement data. The primary goals of this analysis are to (a) demonstrate the estimation of the GP model; (b) provide the first empirical examination to date of how well the data support the assumption of perfect correlation between proximal and future year effects; and (c) examine how inferences about individual teachers from the GP model compare with those from simpler alternatives including the CP and VP models.

3.1. Data

The data used in this analysis are mathematics assessment data from a large urban school district previously explored by Lockwood, McCaffrey, Mariano, and Setodji (2007) in comparison of the VP and CP models. The data contain vertically scaled annual mathematics scores from a national commercial assessment for a cohort of students linked to teachers as they progress from Grade 1 through Grade 5 in academic years 1997–1998 through 2001–2002. Test scores were rescaled so that marginal means ranged from about 3.5 to 6.2 across grades and marginal standard deviations ranged from about 0.92 to 1.08. A total of 9,295 students are present in the data. Only about 21% of students have a complete set of scores for all 5 years. Over all students and grades, student–teacher links were available for only about 54% of records because many students left or entered the district part way through the panel.

Although a vertical scale is present, the mix of topics present on each assessment changes over grades. Certain topics are added or dropped over the assessment sequence and a given topic may not be given equal weight in every year.
Such features are necessarily present in annual assessments that map content to curriculum standards, making this particular data set an interesting and ideal case to study with the relaxed assumptions of the GP model.

3.2. Model Implementation

Under the Bayesian framework, missing test score data are accommodated via the method of data augmentation (Schafer, 1997; Tanner & Wong, 1987; van Dyk & Meng, 2001), which is implemented automatically in WinBUGS under an assumption that the missing scores are missing at random (Little & Rubin, 1987). We handled missing teacher links using the method “M1” of Lockwood, McCaffrey, Mariano, and Setodji (2007), which sets all missing teacher links to a dummy teacher with zero effect.

We used the independent prior distributions described in the Estimation section: normal distributions with mean zero and variance 50 for the marginal means; a Wishart distribution with mean equal to the identity matrix and degrees of freedom equal to $T + 1 = 6$ for $\Sigma^{-1}$; Wishart distributions with degrees of freedom equal to $g + 1$ and means in which $\text{cor}(\theta_{g|t}, \theta_{g|t'}) = 0.8$, for $t \neq t'$, and $\text{var}(\theta_{g|t}) = 0.25 \times 0.7^{(t-g)}$, for $\Gamma_g$, $g = 1, \ldots, 4$; and a uniform distribution (on the range 0 to 0.75) for $\Gamma_5^{1/2}$.

We ran five independent chains of 20,000 iterations each, starting from dispersed initial values of the unknown parameters. Convergence was assessed using the criterion of Gelman and Rubin (1992) and indicated that after 10,000 iterations the chains were sufficiently converged. We saved 10,000 post burn-in iterations from each chain, thinned by 10 for computational efficiency, yielding 5,000 posterior samples for the inferences reported here.

3.3. Results

Posterior means and 95% credible intervals for the diagonal elements (variances) of each $\Gamma_g$ are given in Figure 2. To make the parameters consistently interpretable and comparable regardless of the grade to which they refer, each parameter is expressed as a percentage of the marginal variance of the test score data in the corresponding grade; for example, the Grade 3 elements of $\Gamma_1$, $\Gamma_2$, and $\Gamma_3$ are each scaled relative to the marginal variance of the Grade 3 test score data. The most important result is that the proximal year effects have substantially larger variation than future year effects, suggesting that complete persistence is not appropriate for these data, consistent with the finding of Lockwood, McCaffrey, Mariano, and Setodji (2007). The proximal year variance percentages range from about 14% to more than 35%, while the future year variance percentages are all less than 10% and less than 5% for teachers beyond Grade 1.

Unlike previous analyses of these data, the GP model allows us to test the assumption of perfect correlation among a teacher’s effects from proximal and
future years. Posterior means and 95% credible intervals for the correlation elements of $\Gamma_1$ through $\Gamma_4$ are given in Figure 3. The basic findings are consistent across grades, the proximal year effects are estimated to have correlations about 0.5 to 0.6 with the future year effects, but the future year effects are estimated to have correlations about 0.9 or higher among themselves. Interpreted at face value, these results suggest that the effect that a teacher has on his or her students in the proximal year is of a different nature than the carryover effects to future test administrations, but the future effects are all very similar to one another.

Although we expected the data to have limited information about future year teacher effects, further investigations do not suggest that the estimates for the correlations among the future year effects were unduly influenced by our choice of prior distribution. First, the simulation studies described previously did not suggest that the Wishart prior distributions for the $\{\Gamma_g\}$ were predisposed to produce correlations among the future year effects like those we observed. Second, additional simulations that included cases where students had missing test scores and those scores were more likely to be missing in later years than earlier years again did not find a systematic propensity to produce extremely high correlation.

FIGURE 2. Posterior means (points) and 95% credible intervals (gray lines) for the diagonal elements of each $\Gamma_g$. Each parameter is expressed as a percentage of the marginal variance of the test score data in the corresponding grade.
in later years. Finally, we found the same general patterns using Wishart prior distributions for \( \{ \Psi_g \} \) that set \( \text{cor}(\theta_{g[j][j']}, \theta_{g[j'][j]}) = 0 \).

The finding that the correlation between the proximal and future year effects is less than 1 also does not appear to be unduly influenced by our choice of prior distribution. Our simulation studies indicated that high correlations could be estimated if they truly exist. In addition, the imperfect correlation among the ensemble of effects was replicated using an alternative scale matrices for the Wishart priors for \( \{ \Psi_g \} \) and in a simplification of the GP model where the correlation structure of the \( \{ \Psi_g \} \) matrices was restricted to be compound symmetric. This is the simplest GP generalization of existing models because each \( \Psi_g \) has just one parameter more than the VP model. The results of this model (not shown), which used independent uniform (0,1) prior distributions for the correlation parameters for each grade, had posterior means for the correlations of 0.75, 0.88, 0.71, and 0.69 for Grades 1 to 4, respectively, with upper bounds of 95\% credible intervals of 0.83, 0.98, 0.86, and 0.78. The data thus suggest high but imperfect correlations among the proximal and future year effects and the correlation estimates under the compound symmetric model roughly coincide with a compromise between the moderate and high values reported in Figure 3. However, even with these assurances, we remain concerned that some of the findings about \( \{ \Psi_g \} \), in

![Figure 3](http://jobs.aera.net)

**FIGURE 3.** Posterior means (points) and 95\% credible intervals (gray lines) for the correlations of each \( \Psi_g \) (excluding \( \Psi_5 \), which has no correlation elements).
particular, the high correlation among future year effects, could be driven by an artifact of the missing data patterns or patterns of crossings of students and teachers over time and believe it remains important future work to see whether this finding replicates in other data sets.

By allowing separate effects of each teacher in the proximal and future years, the GP model introduces a large number of nominal parameters, which may lead to overfitting, and it is important to explore whether the additional complexity of the GP model is warranted relative to the simpler CP and VP alternatives. We used the Deviance Information Criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002), a model selection criterion similar to AIC (Akaike, 1973) and BIC (Kass & Raftery, 1995; Schwarz, 1978) but tailored to hierarchical Bayesian models where measuring model complexity is challenging, to examine the value of the additional complexity of the GP model. Like AIC and BIC, DIC uses the sum of two components: model fit, which is small when the model fits the data well, and a penalty for model complexity, which is large for more complex models, to indicate which among a set of candidate models provides the optimal tradeoff between fit and overfit and is best suited to predicting a hypothetical data set generated by the same process as the observed data.

We calculated the model fit and model complexity components of the DIC for the GP, CP, and VP models using the posterior samples described previously. As expected, the model fit component indicated that the GP model provided the best fit, and the model complexity component indicated that it also had the largest effective number of parameters. Combining these two measures led to DIC values of the three models of CP = 8,461, VP = 3,415, and GP = 3,049. Smaller values of DIC indicate preferred models, with differences of 10 or more DIC points generally considered important. Thus, the GP model is clearly preferred to either simpler alternative.

For many purposes, the \( \{ \Gamma_g \} \) may be considered nuisance parameters, with the real focus being on estimates of the teacher effects themselves. And while the GP model provides the ability to estimate the ensemble of proximal and future year effects for teachers, in many practical settings, it might be most important to extract from the GP model a scalar summary of a teacher’s effect on his or her students (e.g., a posterior mean along with a credible interval). Our findings that the future effects are strongly interrelated but are related less strongly to the proximal year effect suggest that single effect may be insufficient and that separate estimates of the proximal year effect and the average effect across all future years may provide a better description of the teachers effects on achievement scores. The question then becomes whether these summaries of teacher effects substantively differ from those of simpler models.

To explore this, we calculated three scalar summaries of teacher effects from the GP model: the posterior mean of the proximal year effect (which we call “GPX”), the posterior mean of the average future year effect (which we call “GPF”), and an equally weighted average of the first two (which we call...
Because of the large differences between the variability of proximal year effects and the variability of future year effects, we rescaled each grade teacher effect by dividing by the square root of the appropriate diagonal element of $\mathbf{H}_g$, expressing all elements of the teacher effect vectors in $z$-score units. We then estimated the posterior mean and standard deviation of the three scalar summaries using the posterior sample draws.

We compared these GP-model estimates to estimates from the CP and VP models. The results are summarized in Table 3, which presents the correlations among the posterior means of these five different scalar summaries of individual teacher effects. In keeping with convention of not focusing on first-year teacher effects because of concerns about selection bias (Ballou et al., 2004; Lockwood, McCaffrey, Mariano, & Setodji, 2007), we present results for Grades 2 through 5 only. Across all grades, we find very high correlation between the proximal year estimates from the GP model and the VP model estimates; these models provide nearly identical estimated teacher effects. The GPX estimates were also very highly correlated (not shown) with estimated proximal year effects from the GP model using the alternative Wishart prior for the $\{\Gamma_g\}$ with 
\[
\text{cor}(\theta_g[p], \theta_g[p']) = 0 \text{ for } t \neq t' \text{ and a Wishart prior distribution for the inverse correlation matrix } \{\Gamma_g^{-1}\}.
\]
As the \( \{\Gamma_g\} \) estimates suggest, the correlation between proximal and future year estimates is only moderate. Estimated teacher effects from the CP model correlate between 0.72 and 0.86 with the proximal year estimates from the GP and VP models, with the largest correlations in Grade 5 where there are no future year effects. These results are consistent with the relationship between teacher estimates from the CP and VP models for these data previously reported by Lockwood, McCaffrey, Mariano, and Setodji (2007). Because the CP model averages students’ future scores with current scores when estimating effects (see Figure 1), the CP model has relatively high correlation with the future year estimates from the GP model. In fact, in Grades 2 and 3, the correlation between the future year and CP estimates is greater than the correlation between the future year and the VP estimates. In addition, the pooled GPA estimates tend to have the largest correlation with estimates from the CP model.

Finally, because the GP model separates the proximal and future year effects, one of its unique benefits is that it can provide information about teachers whose patterns of effects are unusual. For example, we might look for teachers with positive proximal year effects that are substantially larger than their future year effects as potential cases of score inflation (Hamilton et al., 2006). Given the strong correlations among the future year effects estimated in this example, it was natural to compare the average future effect to the proximal year effect (both standardized, as above) in our search for teachers with extreme differences. For each teacher in the data, we calculated the posterior distribution of the GPX minus the GPF. We then determined which differences were detectably different from zero by examining whether the 95\% credible interval for each difference contained zero. This procedure flagged 25 Grade 2 teachers (8\%), 19 Grade 3 teachers (6\%), and 17 Grade 4 teachers (5\%). Interestingly, positive differences (indicating larger proximal year effects than future year effects) outnumbered negative differences by more than 2:1 in each grade. This asymmetry is consistent with the hypothesis that some teachers take actions that inflate students’ current year scores without equal effects to more general measures of achievement.

4. Discussion

As the prospect of using longitudinal achievement data to make potentially high-stakes inferences about individual teachers becomes more of a reality, it is important that statistical methods be flexible enough to account for the complexities of the data. The increasing frequency of tests that are not developmentally scaled across grades, as well as the concerns about the properties of developmental scales, suggests that longitudinal data series may need to be treated as repeated correlated measures of different constructs rather than repeated measures of a consistently defined unidimensional construct. Coupled with the inherent complexity of the accumulation of past educational inputs,
models that assume equality or otherwise perfect correlation between proximal and future year effects of individual teachers may be inappropriate and run the risk of leading to misleading inferences about teachers. The GP model developed in this article tackles these issues head-on by generalizing existing value-added models to handle both scaling inconsistencies across repeated test scores and potential decay in the effects of past educational inputs on future test scores.

The results of our empirical investigations suggest that the assumption of perfect correlation between proximal and future effects of individual teachers is not entirely consistent with the data. This assumption had not been previously examined in the literature, and the GP model provides the ideal framework for carrying out this test. Although the correlation between proximal year and future year effects was only 0.6 and the assumption of perfect correlation of a teacher’s effects is clearly inconsistent with the data, the impact of this assumption on estimated teacher effects for the proximal year was minimal; estimates from the GP model were extremely highly correlated with those from the VP model. For this data set, ignoring the additional complexities of the GP model in favor of the computationally simpler VP model may not be costly for proximal year inference, and the decision between the CP and VP model was of much higher leverage. However, the results here are based on a single data set of assessment scores purported to be developmentally scaled. It is important for future work to carry out similar investigations with other data sets, particularly those with tests that are not on a vertical scale, to understand how generalizable our findings may be.

The additional complexity of the GP model might be useful even if the VP model tends to provide similar estimate of individual teacher effects in the proximal year. As described above, the GP model allows for greater understanding of the relationship among proximal and future year effects and of how teacher effects accumulate. When proximal and future effects differ, important teacher contributions may not be apparent until into the future, as may be the case when teaching advanced materials, or, conversely, having proximal year effects larger then future year effects on the same scale, as was found in our example, raises questions about providing students with a proper foundation in the proximal year.

In this article, we focus on using the GP model for estimating the effects of individual teachers. However, the model can also be used in other applications to account for complex correlation patterns among scores from students who shared classroom assignments. For example, when studying student achievement following an intervention, the GP model could be used to account for students’ multiple classroom assignments and provide accurate standard errors for estimated long-term treatment effects.

The model could be extended to multiple subjects or multiple cohorts of students. Although the necessary adaptations to the model to accommodate multiple subjects or cohorts are straightforward, they are likely to greatly increase the...
computational resources required to estimate the model parameters. In particular, we expect that these changes would limit the utility of WinBUGS for estimating these models, and specialized algorithms like those developed in Lockwood, McCaffrey, Mariano, and Setodji (2007) would likely be required to make estimation feasible. Development of such algorithms can be tedious and may only be worthwhile in applications where there is evidence that the additional flexibility of the GP model versus simpler alternatives may be warranted.

A limitation to the model is the assumption of multivariate normality for the ensembles of teacher effects. Our current formulation assumes that every teacher’s ensemble of effects is drawn from a single distribution and does not allow for groups of teachers to be more likely to have relatively larger or smaller current or future year effects. For example, teachers who narrowly focus on the current test material might have weaker correlation between proximal and future effects. Models that allow for mixtures of latent classes of teachers might capture these differences. If such classes do exist, multivariate normality might not be the best fitting model for the effects. Under certain assumptions about the mixtures, a multivariate t distribution might be better fitting.

Similar to the VP model, the GP model relies on the longitudinal regrouping of students for model identifiability. Insufficient mixing of students’ groupings over time would create a lack of identification of proximal and future effects. A related issue and a potential source of bias of estimated teacher effects is stratification. If the data include sets of teachers that never share any students, then the estimated teacher effects across these students may not be comparable. In our empirical example, we tested for stratification of the teacher and student groups using an algorithm similar to that discussed by Cornelissen (2008) and found that 98% of the teachers and 97% of the students in the data belong to a single connected stratum. Consequently, bias due to stratification is likely to be minimal for this example.

Our model does not explicitly account for the nonrandom assignment of students to classrooms. Models like the persistence models discussed in this article can remove the potential biases from selection into classes, provided there are sufficiently many tests (Lockwood & McCaffrey, 2007). However, it is not clear what effect the more complex teacher effect distribution might have on the estimates compared to the CP and VP models. The increased flexibility for specifying teacher effects might actually degrade the bias reducing advantages of jointly modeling the students’ scores. For example, McCaffrey, Lockwood, Mariano, and Setodji (2005) speculate that the CP model may provide additional safeguards against omitted variable bias compared to the VP model, and it is possible that the GP model might be even more susceptible to such biases than the VP model. The tradeoffs of various models in the presence of selection and misspecification are an area for future research.

Another area for future research is the specification of the prior for $\Gamma_g$. We used an informative Wishart prior for each $\Gamma_g$. This prior was better matched.
to our expectations about the covariance matrices than the more traditional inverse Wishart priors. In addition, the Wishart prior could be implemented in WinBUGS, greatly increasing the applicability of the GP model. Barnard et al. (2000) present an alternative to this prior. These authors suggest decomposing a covariance matrix into variance components and a correlation matrix and propose two classes of distributions for correlation matrix. The first distribution for the correlation matrix is uniform for each correlation coefficient. The second distribution puts equal mass on all conformable correlation matrices. These priors are uninformative in the parameters of interest, which might be desirable for some applications. However, these priors cannot currently be used with WinBUGS and would require users to develop their own MCMC code. Other distributions have been proposed for priors for covariance matrices for multivariate normal data (Daniels, 1999; Daniels & Kass, 1999) and these or other priors should be explored with GP model.

The GP model presents an alternative for modeling longitudinal test score data that accounts for the complex multiple-membership structure of the data and allows for great flexibility when modeling teacher’s effects on students. It provides an alternative for estimating teacher effects and can be fit using available software even for moderately large data sets. Given the increasing desire of policy makers and educators to use varied longitudinal test score data for high-stakes inferences about teachers and other educational inputs, the GP model is one more step toward providing the best statistical tools for these analyses.

Notes

1. As noted in Lockwood, McCaffrey, Mariano, and Setodji (2007), we use the term “teacher effects” when describing the random components included at the classroom level. These effects are not necessarily causal effects or intrinsic characteristics of the teachers; rather, they represent unexplained heterogeneity among students sharing individual teachers. Ideally, they provide information about teacher performance, but there might be many sources of this heterogeneity, including omitted student characteristics or classroom teacher interactions (McCaffrey et al., 2003; Raudenbush, 2004).

2. Rothstein (2008) in the context of estimating teacher fixed effects explores a model that allows teachers to have distinct proximal and future year effects, but he does not consider random effects modeling.

3. Setting $\alpha = 0$ implies that all but the first cell of $\Gamma_g$ would be 0, regardless of the value of $\rho$.

4. Lockwood, McCaffrey, Mariano, and Setodji (2007) explored these data both marginally and jointly with reading assessment data from the same cohort, presenting results from the joint model in their results.

5. Implementing this method in the GP model required a work-around for the fact that matrix nodes in WinBUGS cannot contain a mixture of deterministic
and stochastic elements. Thus, for missing teacher links in Grades 1 through 4, we approximated a dummy teacher with zero effect by using a stochastic node with a prior distribution that was tightly centered on zero. As noted in Lockwood, McCaffrey, Mariano, and Setodji (2007), setting all missing teacher links to a dummy teacher with zero effect may cause the true teacher effects to have a mean other than zero. In the results that follow, the estimated effects of the true teachers have been post hoc centered to have mean zero.

6. Note that Grade 1 teacher effects from these types of models reflect not only the effect of the Grade 1 teacher but also the cumulative effects of any past educational and noneducational inputs to achievement that cluster at the level of the Grade 1 teacher. As such, they are more susceptible to bias by nonrandom assignment of students to teachers (Ballou et al., 2004; Lockwood, McCaffrey, Mariano, & Setodji, 2007) and are often not reported. The fact that both the proximal year and future year variance percentages for Grade 1 teachers are larger than other teachers is consistent with this concern.

References


Cornelissen, T. (2008). The stata command felsdvreg to fit a linear model with two high-dimensional fixed effects. Stata Journal, 8, 170–189.


A Model for Teacher Effects From Longitudinal Data


Authors

LOUIS T. MARIANO is a senior statistician, RAND Corporation, 1200 S. Hayes St., Arlington, VA, 22202; e-mail: Lou_Mariano@rand.org. His research interests include Bayesian hierarchical models, student assessment, teacher accountability, and the evaluation of education reform efforts.

DANIEL F. McCAFFREY is a senior statistician and PNC Chair in Policy Analysis, RAND Corporation, 4570 Fifth Avenue, Suite 600, Pittsburgh, PA, 15213; e-mail: danielm@rand.org. He has published several papers and reports on value-added modeling of teacher effects.

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

Manuscript received April 21, 2008
Revised received November 26, 2008
Accepted May 7, 2009