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

A Correlated Random Effects Model for Nonignorable Missing Data in Value-Added Assessment of Teacher Effects

Andrew T. Karl, Yan Yang and Sharon L. Lohr

JOURNAL OF EDUCATIONAL AND BEHAVIORAL STATISTICS 2013 38: 577
originally published online 3 September 2013
DOI: 10.3102/1076998613494819

Journal of Educational and Behavioral Statistics
2013, Vol. 38, No. 6, pp. 577–603
DOI: 10.3102/1076998613494819

A Correlated Random Effects Model for Nonignorable Missing Data in Value-Added Assessment of Teacher Effects

Andrew T. Karl
Adsurgo LLC

Yan Yang
Arizona State University

Sharon L. Lohr
Westat

Value-added models have been widely used to assess the contributions of individual teachers and schools to students' academic growth based on longitudinal student achievement outcomes. There is concern, however, that ignoring the presence of missing values, which are common in longitudinal studies, can bias teachers' value-added scores. In this article, a flexible correlated random effects model is developed that jointly models the student responses and the student missing data indicators. Both the student responses and the missing data mechanism depend on latent teacher effects as well as latent student effects, and the correlation between the sets of random effects adjusts teachers' value-added scores for informative missing data. The methods are illustrated with data from calculus classes at a large public university and with data from an elementary school district.

Keywords: *generalized linear mixed model; joint model; missing not at random; multiple membership model*

1. Introduction

With increased focus on accountability in education, there is increased interest in measuring teacher and school contributions toward their students' learning. Assessing teachers solely by their current-year students' scores on a standardized test is widely recognized to penalize teachers of disadvantaged students (Braun, 2005); the measures of teacher effectiveness are biased because teacher effects are confounded with their students' characteristics. Value-added models (VAMs) attempt to reduce this bias by estimating the effects teachers have on the academic growth of their students. Rather than simply calculating the average test score for a classroom, as might be done in a naive performance analysis, VAMs control for information on the students' backgrounds, the students' individual test score histories, and contributions of previous teachers to the students' learning. The simplest VAMs use a gain score as a response (Hanushek, 1971) or include the previous year's test score as a covariate in a regression model (Rowan, Correnti, & Miller, 2002); these control for the student's background through the previous year's test score and possibly other covariates. The Colorado growth model (Betebenner, 2009) uses the previous year's test score as a covariate in a quantile regression model. Other VAMs have been defined using mixed models (Ballou, Sanders, & Wright, 2004; Harris and McCaffrey, 2010; Lockwood, McCaffrey, Mariano, & Setodji, 2007; Mariano, McCaffrey, & Lockwood, 2010; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Raudenbush & Bryk, 2002; Sanders, Saxton, & Horn, 1997; Wright, White, & Sanders, 2010), in which the response is a vector of student scores over time and teacher contributions are modeled through random effects. In mixed models, the empirical best linear unbiased predictors (EBLUPs) of the teacher random effects serve as the VAM scores. As noted by Lockwood et al. (2007), these EBLUPs summarize unexplained heterogeneity at the classroom level, though they are often referred to as *teacher effects*.

VAM scores may be used for a variety of purposes, from identifying needs for professional development to high-stakes purposes such as promoting or firing teachers or closing schools. Many researchers and policymakers have expressed concern about whether VAM scores have sufficient accuracy for high-stakes purposes (Baker et al., 2010; Braun, Chudowsky, & Koenig, 2010; Briggs & Domingue, 2011; Harris, 2011; The National Academies, 2009). If the model assumptions are met and the model contains all relevant information, the VAM scores from that model will be unbiased estimates of the teacher effects on the responses measured. The model assumptions are strong, however, and there is concern about how often some of the assumptions are met in practice. The models assume that students are assigned to teachers randomly or in a noninformative manner (Rothstein, 2010), that the responses are valid measures of student achievement (Koretz, 2008), and that all relevant information is captured in the model. Lohr (2012) discusses the assumptions of various models and shows how violations of the assumptions may be used to manipulate VAM scores.

The models cited above also assume that every student has complete data over the time period studied or that missing data patterns have no information about teacher effectiveness. Ballou, Sanders, and Wright (2004) note that longitudinal mixed model approaches allow students to have missing test scores for some years by including a partial vector of responses, but such analyses assume that the data are missing at random (MAR). Missing data are ubiquitous in longitudinal education data. Students drop courses, change schools, move away, or may be absent on the day of a test. Inference based on analyses of data where some observations are missing requires assumptions about the nature of the missing data. In the college setting, students in Calculus 2 who do not finish Calculus 3 will have missing data for Calculus 3. The missingness may be relevant to estimates of the Calculus 2 teachers' contributions. A student who is poorly prepared for Calculus 3 may drop the class despite having received a high grade in Calculus 2. Or, in the elementary or secondary school settings, it is possible that low-performing students might be discouraged from taking a standardized exam (Fernandez, 2012; Ryan & Weinstein, 2009). In a simplistic example, suppose that students are randomized to one of the several classrooms and a gain score model is used. If teachers were to discourage their weakest students from taking the exam, they could inflate their class averages, and thus their rankings.

The assumptions about missing data made by VAMs have been recognized as a potential problem for their use in teacher evaluation (Amrein-Beardsley, 2008; Braun, 2005; McCaffrey, Lockwood, Koretz, & Hamilton, 2003). McCaffrey, Lockwood, Mariano, and Setodji (2005) and Wright (2004) explore the impact of the presence of missing data on VAMs, though they do not perform a joint analysis of the test scores and missing data indicators. To date, the only thorough investigation of the impact of missing data on VAMs by jointly modeling the test scores and missing data process comes from McCaffrey and Lockwood (2011). They use selection and pattern-mixture models for the missing data indicators with Bayesian inference, attributing attendance to intrinsic student—but not teacher—characteristics.

In this article, we develop a new multiple response, multiple membership mixed model that allows the missing data mechanism to depend on teachers as well as students. This model allows detection of teachers' possible effects on their students' future course taking or their students' attendance during an exam. Because the true responses are missing, the model cannot be used to say for certain that teacher VAM scores would change if the missing data were taken into account, but the model in this article allows exploration of possible effects of missing data on the teacher rankings through a sensitivity analysis (Xu & Blozis, 2011). If the rankings of teacher effects change depending on the assumptions made about the structure of the missing data mechanism, then the possible dependence on missing data should be considered when contemplating high-stakes usages of VAM scores. Even if the teacher effects do not show sensitivity to the structure of the missing data mechanism, the model may be useful as a diagnostic tool. In some situations, no relationship would be expected between the teacher effects and the corresponding effects in the missing data mechanism. By fitting the model and examining a scatter plot of the effects, unusual cases may be discovered.

The article is organized as follows. Section 2 of this article presents background on missing data analyses and the framework for modeling the test scores and the missing data mechanism jointly. Section 3 applies the joint model to calculus data from a large public university. Structures available within the model are used to perform a sensitivity analysis on the teacher rankings produced when analyzing a data set containing semester calculus grades. Section 4 summarizes the results of the model when applied to elementary school math scores. Finally, Section 5 discusses implications of model estimates for uses of VAMs and other applications in which the models developed in this article can describe potential effects of missing data.

2. A Correlated Random Effects Model

Let y_{ig} be the potential response (often, a test score) of student i at time g, for $i = 1 \ldots n$ and $g = 1, \ldots, T$, with $\boldsymbol{y_i} = (y_{i1}, \ldots, y_{iT})'$ and $\boldsymbol{y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_n')'$. The indicator variable

$$r_{ig} = \begin{cases} 1 & \text{if } y_{ig} \text{ is observed} \\ 0 & \text{otherwise} \end{cases}$$

tracks whether the planned measurement on student i at time g is observed or missing. Let $\mathbf{r}_i = (r_{i1}, \ldots, r_{iT})'$ and $\mathbf{r} = (\mathbf{r}_1', \ldots, \mathbf{r}_n')'$. The complete data vector $\mathbf{y} = \{\mathbf{y}^o, \mathbf{y}^m\}$ consists of both the observed data $\mathbf{y}^o$ and the missing data $\mathbf{y}^m$. The vector $\mathbf{y}^o$ consists of the values y_{ig} such that $r_{ig} = 1$, and $\mathbf{y}^m$ consists of the values y_{ig} that would have been observed if the observations were not missing. Since $r_{ig} = 1$ if we observe the value y_{ig}, we refer to the model generating the r_{ig} as the attendance process, where by "attendance" in a particular year we simply mean that a student has a test score recorded for that year. We refer to the model generating the scores y_{ig} as the longitudinal or the score process.

Data may be missing from a study for several reasons, and the cause of the missingness determines the degree to which the missing data affect the analysis. If data are missing completely at random, then the joint likelihood of the longitudinal and attendance processes factors cleanly, and there is no need for joint modeling, since the longitudinal and attendance processes are independent. Likewise, if the data are MAR and the parameters for the longitudinal and missingness processes are distinct, then the missing data mechanism is said to be ignorable for likelihood inference (Little & Rubin, 2002). However, if the missing data are missing not at random (MNAR) and hence nonignorable, then the longitudinal and missingness processes cannot be factored in the likelihood; they must be modeled jointly to explore the effects of missingness on estimates in the longitudinal process.

McCaffrey and Lockwood (2011) have developed selection and pattern-mixture VAMs for nonignorable missing data in which the missing data mechanism depends on latent effects of the students. We expand the availability of VAMs for data with potentially nonignorable missing data by presenting a correlated parameter model (CPM), a generalization of a shared parameter model (SPM: Wu & Carroll, 1988). In the CPM, random effects are included for the latent teacher and student effects in the longitudinal model, a different set of random effects are included for the latent teacher and student effects in the attendance model, and the two sets of random effects are allowed to be correlated (Lin, Liu, & Zhou, 2009). Allowing correlated rather than shared random effects as in the SPM avoids the SPM's restriction that the random effects have the same variance and structure. The CPM proposed in this article allows the missing data mechanism to depend on the effects of teachers as well as students. This gives more flexibility in detecting sensitivity to missing data, since it is plausible that the missing data trajectory of students could depend on their current and former teachers.

The CPM produces the observed data likelihood via the factorization

$$f(\mathbf{y^o}, \mathbf{r}) = \iint f(\mathbf{y^o}|\boldsymbol{\eta}_{\text{score}})\, f(\mathbf{r}|\boldsymbol{\eta}_{\text{attnd}})\, f(\boldsymbol{\eta}_{\text{score}}, \boldsymbol{\eta}_{\text{attnd}}) d\boldsymbol{\eta}_{\text{score}} d\boldsymbol{\eta}_{\text{attnd}}, \tag{1}$$

where $f(\boldsymbol{\eta}_{\text{score}}, \boldsymbol{\eta}_{\text{attnd}})$ is the density of a multivariate normal distribution. The vector $\boldsymbol{\eta}_{\text{score}}$ contains random student and teacher intercepts for the longitudinal process, while the vector $\boldsymbol{\eta}_{\text{attnd}}$ contains a flexible combination of student and/or teacher effects for the attendance mechanism. The CPM assumes that the longitudinal and attendance processes are conditionally independent, given the random effects.

CPMs make different assumptions on the joint model than selection and pattern-mixture models (e.g., conditional independence [CI]) and present an alternative approach for missing data modeling. The CPM framework allows for straightforward inclusion of teacher history in the modeling of the drop-out mechanism. The EBLUPs of the classroom effects in the attendance model provide a direct method of evaluating the frequency with which teachers' students have missing data. Since the attendance model estimates the probability that a given observation would be recorded, a larger EBLUP for a classroom effect in the attendance model indicates that students who took that particular class are more likely to complete the next year than students who took another class that year (i.e., with another teacher). It would, however, be unrealistic to expect the effect of a teacher on student learning to be identical to the effect of the teacher on the future student attendance, so $\boldsymbol{\eta}_{\text{score}}$ and $\boldsymbol{\eta}_{\text{attnd}}$ are assumed to be correlated rather than identical.

2.1. The Observed Data Model

We now present the model $f(\mathbf{y}^o|\boldsymbol{\eta}_{\text{score}})$ for student scores $\mathbf{y}^o$ using information about the history of observations on each student and each student's teacher history. We use the generalized persistence (GP) model of Mariano et al. (2010) for the longitudinal mechanism. The GP model is among the most general of the mixed models used for VAMs and contains many of the other mixed models as special cases. If the data are MAR, the model in Equation 1 reduces to the

GP model. Suppose a data set tracks a cohort of n students over T years. The GP model assumes a linear mixed model as follows:

$$y^o_{ig} = \boldsymbol{x}'_{ig}\boldsymbol{\beta}_{\text{score}} + \mathbf{s}'_{ig}\boldsymbol{\eta}_{\text{score}} + \varepsilon_{ig}, \tag{2}$$

where y^o_{ig} denotes the score for student i during year g, for $i = 1, \ldots, n$, and $g \in A_i$; A_i is the set of years in which student i is observed. Students are taught by one of m_g teachers in each year g. We will also refer to the vector of concatenated student scores, $\boldsymbol{y}^o = \left(\boldsymbol{y}^{o\prime}_1, \ldots, \boldsymbol{y}^{o\prime}_n\right)'$, where $\boldsymbol{y}^{\mathbf{o}}_{\mathbf{i}} = \left(y^o_{ig}\right)$. The matrix $\mathbf{X}$, with rows $\boldsymbol{x}'_{ig}$, is the design matrix for the vector $\boldsymbol{\beta}_{\text{score}}$ of student- and teacher-level covariates such as demographic information or years of teaching experience. The matrix $\mathbf{S}$, with rows $\mathbf{s}'_{ig}$, indicates which students and teachers are associated with the responses in $\mathbf{y}^o$.

The random effects vector $\boldsymbol{\eta}_{\text{score}} = \left[\delta'_{score}\theta'_{score}\right]'$ has two components. Student i has a latent effect δ_i that represents an underlying level of achievement not explained by the fixed covariates, and $\boldsymbol{\delta}'_{\text{score}} = (\delta_1, \ldots, \delta_n)'$. We assume that $\delta_1, \ldots, \delta_n$ are independent and identically distributed $N_1(0, \Gamma_{\text{stu}})$ random variables. This represents a slight departure from Mariano et al. (2010), who model the intrastudent correlation in an unstructured error covariance matrix. However, that structure is not as amenable to the joint model for attendance because it precludes the possibility of including student effects in the attendance model. As a result, we model the intra-student correlation with random effects, similar to the VAM used by McCaffrey and Lockwood (2011). When the responses y_{ig} all have the same scale for $g = 1, \ldots, T$, this leads to a compound symmetry covariance structure for the students. If the student random effects are omitted from the missing data mechanism, the intra-student correlation may be modeled in an unstructured error covariance matrix as done by Mariano et al. (2010).

The GP model estimates the effect of teachers on students in the year that they teach them, their lasting effect on the next year's score, and so on. Following the notation of Mariano et al. (2010), we let $\theta_{g[jt]}$ represent the effect for the jth grade-g teacher on a student's grade t score. A grade $g = 1, \ldots, T$ teacher has $K_g = T - g + 1$ effects. Thus, $\boldsymbol{\theta}_{g[j\cdot]}$ gives the vector of current and future year effects of the jth grade g teacher. The vector $\boldsymbol{\theta}_{\text{score}}$ concatenates the $\boldsymbol{\theta}_{g[j\cdot]}$ effects for all grades and teachers. The model is able to distinguish between the persistence effect of former teachers and the current effect of the present teacher because the students are not nested at the teacher level. The design matrix $\mathbf{S}$ of the random effects has rows s'_{ig}, and may be partitioned into two blocks $\mathbf{S} = [\mathbf{S}_1\ \mathbf{S}_2]$. $\mathbf{S}_1$ contains a 1 in column i if the observation is for student i, and $\mathbf{S}_2$ contains 1's in entries corresponding to teachers who could affect that response. We specify the structure and distribution of the random effects in Section 2.3.

The error terms are distributed as $\boldsymbol{\varepsilon} \sim N(0, \mathbf{R})$ where $\mathbf{R}$ is a diagonal matrix with entries coming from the set $\{\sigma_1^2, \ldots, \sigma_T^2\}$, depending on the year of the observation. In addition, we assume $\mathrm{cov}(\boldsymbol{\eta}_{\mathrm{score}}, \boldsymbol{\varepsilon}) = 0$.

2.2. The Attendance Model

In the attendance model, the probability p_{ig} that student i provides a score at time g (i.e., $r_{ig} = 1$) depends on covariates and latent teacher and/or student effects. We use a threshold model for p_{ig}, the conditional probability that $r_{ig} = 1$ (McCulloch, 1994). Using a probit link, the generalized linear mixed model (GLMM) is

$$\begin{aligned} r_{ig} | \boldsymbol{\eta}_{\mathrm{attnd}} &\sim \mathrm{Bin}(1, p_{ig}) \\ \Phi^{-1}(p_{ig}) &= \boldsymbol{w}'_{ig} \boldsymbol{\beta}_{\mathrm{attnd}} + \boldsymbol{z}'_{ig} \boldsymbol{\eta}_{\mathrm{attnd}}. \end{aligned}$$

The vectors $\boldsymbol{w}'_{ig}$ and $\boldsymbol{z}'_{ig}$ describe which fixed and random effects are thought to be related to the response mechanism. The vector of fixed effects $\boldsymbol{\beta}_{\mathrm{attnd}}$ of the attendance model will be different from the $\boldsymbol{\beta}_{\mathrm{score}}$ of the observed model. It will represent a baseline propensity for attendance at each level of the fixed effects. Furthermore, the attendance model requires that there is at least one missing observation at each level of each categorical fixed effect in the attendance mechanism. Otherwise, the data suffer from quasi-complete separation (Allison, 2008). In that case, the maximum likelihood estimate for the particular fixed effect does not exist.

We may include either random teacher effects, random student effects, or both in $\boldsymbol{\eta}_{\mathrm{attnd}}$. The structure of the random effects is flexible and may be modified depending on the goals of the study. This flexibility provides the means for performing a sensitivity analysis. When jointly modeling MNAR data, the CPM makes untestable assumptions about the nature of the relationship between the observed data and attendance processes. Molenberghs, Kenward, Verbeke, Beunckens, and Sotto (2008) show that it is not possible to perform an overall test of MNAR versus MAR since every MNAR model has an MAR counterpart that provides the same fit to the observed data but different predictions for the unobserved data. The plausibility of the assumed model cannot be tested empirically, and as a result it is necessary to fit several alternatives of the attendance model to check the sensitivity of the inference to the choice of joint modeling structure (Xu & Blozis, 2011).

The student effects in the attendance model, if included, will be denoted by $\delta_i^{\mathrm{attnd}}$. The teacher effects in the attendance model will be denoted by $\Lambda_{g[j]}$. These effects may be structured in a number of different ways. In our application in Section 3, $\Lambda_{g[j]}$ represents the effect that the jth grade g teacher has on the probability of his or her students being measured in year $g + 1$. This effect measures how likely it is that students are observed in the year after studying under a particular teacher. This effect is not calculated for teachers in the last year of observations (year T) because no information is available on the future drop-out patterns of students of those teachers. This feature of the model would detect instructors whose students drop out (of the school or sequence of courses) at a relatively high rate. We refer to these effects as the "attendance effects" of the grade g teachers, since they measure the rate with which students complete year $g + 1$. This models the effects of teachers on their students' future course taking as well as on their completion of subsequent courses.

In other settings, it makes more sense to model the effect of missing data in the current year, g. For example, in the grade school application in Section 4, we structure the missing data mechanism to measure the proportion of each grade-g teacher's students who actually take the standardized exam in that year. In the calculus example, we may wish to distinguish between students who drop out of a Calculus 3 course and those who never enrolled. If information about students who drop courses is available, it would be reasonable to use the attendance effect of a grade-g teacher to model the proportion of students who complete their course. The model is flexible and allows for many variations on the implementation of the missing data mechanism. The attendance mechanism may be used to model the effects of year g teachers on attendance in year g, on attendance in year $g + 1$, or on both, assuming different random effects for the 2 years. When both teacher and student effects are included in the attendance model, it is important to make sure those effects are defined to model the same concept.

The conditional density of r_{ig} given the random effects vector $\boldsymbol{\eta}_{\text{attnd}}$ (which contains the effects δ_i^{attnd} and $\Lambda_{g[j]}$) is

$$f\left(r_{ig}\middle|\boldsymbol{\eta}_{\text{attnd}}\right) = \Phi\left((-1)^{1-r_{ig}}\left[\boldsymbol{w}'_{ig}\boldsymbol{\beta}_{\text{attnd}} + \boldsymbol{z}'_{ig}\boldsymbol{\eta}_{\text{attnd}}\right]\right).$$

As with the y_{ig}, we assume the r_{ig} are conditionally independent given the random effects, yielding

$$f(\mathbf{r}|\boldsymbol{\eta}_{\text{attnd}}) = \prod_{i=1}^{n}\prod_{g=1}^{T}\Phi\left((-1)^{1-r_{ig}}\left[\boldsymbol{w}'_{ig}\boldsymbol{\beta}_{\text{attnd}} + \boldsymbol{z}'_{ig}\boldsymbol{\eta}_{\text{attnd}}\right]\right).$$

2.3. The Joint Model

In typical usage, VAMs assume that data are MAR. Inference is intended to be on $\mathbf{y} = (\mathbf{y}^o, \mathbf{y}^m)$, but only the $\mathbf{y}^o$ have been observed. When data are MNAR, $f(\mathbf{y}^o)$ is not the correct likelihood to maximize because $\mathbf{r}$ provides information about the distribution of $\mathbf{y}$. As a result, the longitudinal and attendance processes must be modeled jointly and $f(\mathbf{y}^o, \mathbf{r})$ must be maximized. We construct the joint model via the correlated random effects factorization (Equation 1). The formulation presented in this section assumes attendance in year $g + 1$ is being modeled.

We concatenate the random effects vectors $\boldsymbol{\eta}_{\text{score}}$ and $\boldsymbol{\eta}_{\text{attnd}}$ into a single random effects vector, $\boldsymbol{\eta}$. To ensure that the $\text{cov}(\boldsymbol{\eta}) = \mathbf{G}$ matrix is block diagonal, we structure the $\boldsymbol{\eta}$ vector as

$$\begin{aligned}\boldsymbol{\eta} = (&\delta_1, \delta_1^{\text{attnd}}, \ldots, \delta_n, \delta_n^{\text{attnd}}, \boldsymbol{\theta}_{\mathbf{1}[\mathbf{1}\cdot]}, \Lambda_{1[1]}, \ldots, \boldsymbol{\theta}_{\mathbf{1}[m_{\mathbf{1}}\cdot]}, \Lambda_{1[m_1]}, \boldsymbol{\theta}_{\mathbf{2}[\mathbf{1}\cdot]}, \\ &\Lambda_{2[1]}, \ldots, \boldsymbol{\theta}_{\mathbf{2}[m_{\mathbf{2}}\cdot]}, \Lambda_{2[m_2]}, \ldots, \boldsymbol{\theta}_{\mathbf{T}[m_{\mathbf{T}}\cdot]})'.\end{aligned} \tag{3}$$

We model the random student effects and their counterparts for the attendance mechanism, if they are included, as $\left(\delta_i, \delta_i^{\text{attnd}}\right)' \sim N_2(\mathbf{0}, \Gamma_{\text{stu}})$ where Γ_{stu} is a 2×2 unstructured covariance matrix. If the random student effects are not included in the attendance model, simply omit the δ_i^{attnd} from $\boldsymbol{\eta}$ and model $\delta_i \sim N_1(0, \Gamma_{\text{stu}})$. The teacher effects are assumed independent of the student effects and distributed as

$$\begin{cases} \left(\boldsymbol{\theta}'_{g[j\cdot]}, \Lambda'_{g[j]}\right)' \sim N_{K_g+1}\left(0, \Gamma_g\right) & \text{if } g \neq T \\ \left(\boldsymbol{\theta}'_{g[j\cdot]}\right)' \sim N_{K_g}\left(0, \Gamma_g\right) & \text{if } g = T \end{cases},$$

where Γ_g is an unstructured covariance matrix. Then

$$\mathbf{G} = \text{cov}(\eta) = \text{blockdiag}(\Gamma_{\text{stu}}, \ldots, \Gamma_{\text{stu}}, \Gamma_1, \ldots, \Gamma_1, \ldots, \Gamma_T, \ldots, \Gamma_T), \tag{4}$$

where there are n copies of Γ_{stu}, and for each $g = 1, \ldots, T$ there are m_g copies of Γ_g in $\mathbf{G}$. The $\mathbf{R}$ matrix for $f(\mathbf{y}^o|\boldsymbol{\eta}_{\text{score}})$ is unchanged from Section 2.1. The log likelihood for the joint model in Equation 1 may be expressed as

$$l(\boldsymbol{\Psi}) = \log \iint \prod_{i=1}^{n} \left\{ \prod_{g \in A_i} f\left(y_{ig}^o \middle| \boldsymbol{\eta}_{\text{score}}\right) \prod_{g=1}^{T} f\left(r_{ig} \middle| \boldsymbol{\eta}_{\text{attnd}}\right) \right\} f(\boldsymbol{\eta}_{\text{score}}, \boldsymbol{\eta}_{\text{attnd}}) d\boldsymbol{\eta}_{\text{score}} d\boldsymbol{\eta}_{\text{attnd}}, \tag{5}$$

where

$$\begin{aligned} f\left(y_{ig}^o \middle| \boldsymbol{\eta}_{\text{score}}\right) &\propto \left(\sigma_g^2\right)^{-1/2} \exp\left[-\left(y_{ig}^o - \mathbf{x}'_{ig}\boldsymbol{\beta}_{\text{score}} - \mathbf{s}'_{ig}\boldsymbol{\eta}_{\text{score}}\right)^2 / \left(2\sigma_g^2\right)\right], \\ f\left(r_{ig} \middle| \boldsymbol{\eta}_{\text{attnd}}\right) &= \Phi\left[(-1)^{1-r_{ig}}\left(\mathbf{w}'_{ig}\boldsymbol{\beta}_{\text{attnd}} + \mathbf{z}'_{ig}\boldsymbol{\eta}_{\text{attnd}}\right)\right], \\ f(\boldsymbol{\eta}_{\text{score}}, \boldsymbol{\eta}_{\text{attnd}}) &= f(\boldsymbol{\eta}) \propto \det(\mathbf{G})^{-1/2} \exp\left[-\left(\boldsymbol{\eta}'\mathbf{G}^{-1}\boldsymbol{\eta}\right)/2\right]. \end{aligned}$$

A_i is the set of years in which student i has an observation, and $\boldsymbol{\Psi}$ is a vector of the model parameters.

Note that the models are specified separately: The model of the test scores y_{ig} contains only the parameters $\boldsymbol{\beta}_{\text{score}}$ and the random effects δ_i and $\boldsymbol{\eta}_{\text{score}}$; the model of the attendance indicators r_{ig} contains only the parameters $\boldsymbol{\beta}_{\text{attnd}}$ and the random effects δ_i^{attnd} and $\boldsymbol{\eta}_{\text{attnd}}$. The effects $\boldsymbol{\eta}_{\text{score}}$ and $\boldsymbol{\eta}_{\text{attnd}}$ are related through the correlation structure in the matrix $\mathbf{G}$. If student i is absent at time g, there will be no observation for y_{ig}, but $r_{ig} = 0$ will still be modeled: The correlation between the random effects in the two models means that the missing value contributes to the estimates of student and teacher effects in the test score model.

2.4. Estimation

The joint model presents a high-dimensional integration problem when calculating the marginal distribution of the observed data in Equation 5. The source of the problem is twofold, due to the presence of a nonlinear link in the integrand for the modeling of the binary attendance process and the multiple membership structure of VAMs. The random effects' correlation structure is not nested, which means that the integral over the random effects cannot be factored into a product of low-dimensional integrals (e.g., one- or two-dimensional integrals). Even under the assumption of MAR and without the integration problem, the GP model is computationally demanding because of its random effects structure. Mariano et al. (2010) notice sensitivity to the choice of prior distributions for the covariance matrices when estimating the GP model with Bayesian methods. Karl, Yang, and Lohr (2013b) use an expectation–maximization (EM) algorithm to develop an efficient maximum likelihood routine for estimating the GP model (Mariano et al., 2010) under an assumption of MAR. The EM algorithm is available through the R (R Core Team, 2013) package GPvam (Karl, Yang, & Lohr, 2012). The general method for estimating the parameters of nonnested, multiple response GLMMs developed in Karl, Yang, and Lohr (2013a) is used to perform calculations for the CPM in this article. This method makes use of first-order and fully exponential Laplace approximations for the intractable integrals that appear in the E step (Rizopoulos, Verbeke, & Lesaffre, 2009; Steele, 1996).

3. Effects of Missing Data in Calculus Classes

This section applies the model to data on calculus grades from a large public university. Broatch and Lohr (2012) use a subset of these data in their analyses. The data set tracks 3,557 students who took Calculus 2 and possibly Calculus 3 at the university. A total of 184 Calculus 2 classes are included from Fall 2000 through Spring 2005. In addition, 144 Calculus 3 classes from Spring of 2001 through Spring of 2006 are included. Students who took only Calculus 3 during the study are omitted. Each classroom is treated as a separate effect. Effects corresponding to different classes taught by the same teacher are assumed to be independent. An alternative model could be fit in which classes taught by the same teacher are nested within that teacher and an additional random effect added at the teacher level. In that case, it would be expected that the mean responses of classes taught by the same teacher would be positively correlated. Accounting for this correlation would result in slightly larger standard errors for the estimated teacher effects. Another approach would be to introduce additional parameters to the appropriate off block diagonal components of **G**, explicitly modeling the correlation between classroom effects belonging to the same teacher.

Analysis focuses on the grades assigned to students, which are converted to the corresponding value on a 4-point scale. The scores in the data set are collectively centered and standardized. With $+/-$ grades, there are eight possible numeric values for the student scores. The normal approximation for the error terms seems reasonable, though the quality of the approximation would deteriorate as the number of distinct grades decreases.

3.1. Sensitivity Analysis

In this data set, only 2,140 of the 3,557 students who completed Calculus 2 also completed Calculus 3. Longitudinal sequences in the university setting often have a different pattern of missing data than longitudinal data sets in the elementary school setting, because missing data in universities are often due to students' decisions to drop out of college, to change majors, or simply not to complete the calculus sequence. These decisions may be influenced by the students' previous or current instructors. In the models shown here, the attendance variable for Calculus 3 is modeled as a function of the effect of the Calculus 2 instructor. Some students may have such a poor experience with a particular instructor that they decide to not to take the next course in the sequence, or upon beginning the next course find themselves unprepared and drop out. Of course, a student's completion of Calculus 3 is a function of many other things besides his or her experience with his or her Calculus 2 instructor.

Our goal is not to select a particular attendance mechanism, but rather to test the sensitivity of teacher EBLUPs to assumptions about missing observations. As Molenberghs and Kenward (2007) discuss, focusing attention on one particular MNAR model is no better than ignoring MNAR models. The observed data cannot provide evidence for or against the MAR assumption without an a priori assumption about the correct form of the MNAR model (Rhoads, 2012). The choice of attendance mechanism must be made from a subject–matter perspective. When an alternate attendance mechanism provides a plausible representation of the missing data process and yields substantially different teacher effects from the test score model, then the accuracy of the MAR rankings is questionable. While the lack of sensitivity of the EBLUPs to different choices of attendance mechanisms strengthens our confidence in the results, it is always possible that the missing observations are nonignorable according to an untested attendance mechanism.

We fit a model using just the yearly means as fixed effects in both the score and the attendance models (Model 1), as well as a model that includes gender, race/ethnicity, and SAT quantitative score (SATQ) as covariates in both the score and attendance models (Model 2). Because some of the students do not take the SAT, we treat SATQ as a categorical variable with six categories: the five quintiles of scores, with a sixth category for students who did not take the SAT. Because the student scores come from nonstandardized class grades, the current-year teacher effects reflect the tendency of individual teachers to assign above- or below-average grades, and not necessarily the effectiveness of their teaching. The future year effects of Calculus 2 teachers, however, reflect how well each teacher's former students performed in comparison to their new Calculus 3 classmates. Our investigation focuses on these future year effects.

While not every student who takes Calculus 2 does so with the intention of taking Calculus 3, we may expect to see, on average, a certain proportion of Calculus 2 students going on to complete Calculus 3. In this example, we construct the attendance mechanism to measure the proportion of students from Calculus 2 classes who complete Calculus 3. To perform a sensitivity analysis, we fit an MAR model and compare its estimated teacher effects to those from three different MNAR models.

In the model we call MNAR-t, we include a random teacher effect for Calculus 2 teachers in the attendance mechanism that is correlated with the corresponding teacher effects from the observed data mechanism and measures the proportion of each teacher's students who go on to complete Calculus 3. The model MNAR-s models Calculus 3 completion as a function of student random effects. Even though only one binary observation is made on each student, we are able to fit this model because the predicted student effects in the attendance mechanism borrow strength from their correlation with the student effects from the observed data mechanism. Finally, MNAR-b contains both random student and teacher effects in the attendance mechanism. The appropriate attendance process cannot be chosen by empirical investigation of the observed data (including examination of the log likelihood) since the observed data do not provide information to support one particular MNAR model over another (Fitzmaurice, Laird, & Ware, 2004; Xu & Blozis, 2011). Instead, we compare the estimated teacher effects across different models, looking for sensitivity to the assumptions about the nature of the missing data.

3.2. Results

The parameter estimates for Model 1 appear in Table 1. The covariance parameter estimates for Model 2 are very similar. The estimates for the fixed effects of Model 2 appear in Table 2. The yearly means in the observed data model are represented by μ_i^y, for $i = 1, 2$. The value μ_2^r gives the estimated proportion, for example, $\Phi(0.246) = 0.597$, of Calculus 2 students who complete Calculus 3. The other parameters follow the same notation as used in Section 2. Also listed for each model are -2 times the Laplace-approximated log likelihood ($-2l$) and the correlation (ρ) of the predicted Calculus 2 future year effects with those from the MAR model. This correlation provides a summary of the sensitivity of the teacher rankings to assumptions about the nature of student dropout under different models for the attendance mechanism. Using selection and pattern-mixture models to model the drop-out process as a function of student effects, McCaffrey

TABLE 1
Sensitivity Analysis for Model 1

	MAR	MNAR-t	MNAR-s	MNAR-b
μ_1^y	−0.095 (0.027)	−0.097 (0.028)	−0.092 (0.027)	−0.094 (0.028)
μ_2^y	−0.154 (0.034)	−0.161 (0.035)	−0.282 (0.035)	−0.284 (0.035)
μ_2^r		0.246 (0.026)	0.307 (0.065)	0.304 (0.041)
σ_1^2	0.388 (0.023)	0.385 (0.023)	0.328 (0.020)	0.330 (0.020)
σ_2^2	0.292 (0.019)	0.293 (0.019)	0.330 (0.019)	0.329 (0.019)
$\Gamma_{\text{stu}}[1,1]$	0.618 (0.026)	0.620 (0.026)	0.680 (0.026)	0.674 (0.025)
$\Gamma_{\text{stu}}[2,1]$			0.637 (0.128)	0.640 (0.065)
$\Gamma_{\text{stu}}[2,2]$			0.600 (0.633)	0.610 (0.261)
$\Gamma_1[1,1]$	0.082 (0.015)	0.085 (0.015)	0.077 (0.013)	0.082 (0.015)
$\Gamma_1[2,1]$	−0.004 (0.009)	−0.001 (0.010)	−0.006 (0.009)	−0.002 (0.010)
$\Gamma_1[3,1]$		0.044 (0.015)		0.017 (0.013)
$\Gamma_1[2,2]$	0.028 (0.011)	0.031 (0.011)	0.028 (0.010)	0.030 (0.011)
$\Gamma_1[3,2]$		0.021 (0.009)		0.010 (0.012)
$\Gamma_1[3,3]$		0.040 (0.014)		0.052 (0.022)
Γ_2	0.080 (0.015)	0.082 (0.015)	0.082 (0.015)	0.082 (0.015)
$-2l$		20,022.7	19,447.6	19,436.7
ρ	1	0.881	0.994	0.984

Note: MAR = missing at random; MNAR = missing not at random.
Standard errors are in parentheses.

and Lockwood (2011) found values of ρ that were all greater than 0.97. MNAR-s provides the analog of their models using correlated random effects, and yields $\rho = 0.994$. Likewise, MNAR-b does not produce teacher effects that are substantially different from the MAR model. However, MNAR-t reorders the teacher effects, producing $\rho = 0.881$.

Aaronson, Barrow, and Sander (2007) rank teachers by the quartile of the relevant effect that their individual estimate falls in. While sometimes used in practice for personnel decisions, a simple division of the classrooms into quartiles does not account for the error in the estimates of the classroom effects. Analyzing the calculus data with MNAR-t leads to different classifications with the quartile method than those produced by MAR model. Thus, a teacher may receive a different evaluation based on the model assumed (either tacitly or explicitly) for the attendance mechanism. Using the method of Aaronson et al. (2007), some teachers move two (or even three) quartiles when evaluated with MNAR-t, as shown in Table 3. Figure 1 plots the Calculus 2 future year teacher effects from MNAR-t against the future effects from the MAR model. The quartile ranking appears to be relatively sensitive to the assumed nature of the missing data, although the confidence intervals for estimated teacher effects may also be wide. Of the 83 classrooms that change quartiles, 73 of those change only one quartile. These changes could be as simple as, for example, a shift from the 26th to the 24th percentile.

TABLE 2
Fixed-Effects Estimates for Model 2 Assuming MNAR-t

	$f(\mathbf{y})$		$f(\mathbf{r})$	
μ_1^y	.602	(.069)	—	—
μ_2^y and μ_2^r	.539	(.071)	.459	(.096)
Female	—	—	—	—
Male	−.155	(.035)	.119	(.049)
Asian	—	—	—	—
Black	−.603	(.104)	−.315	(.147)
Hispanic	−.231	(.065)	−.203	(.094)
Native American	−.662	(.111)	−.375	(.156)
Missing race	.088	(.071)	.126	(.106)
White	−.198	(.049)	−.199	(.072)
SATQ-5	—	—	—	—
SATQ-4	−.140	(.058)	−.023	(.086)
SATQ-3	−.378	(.057)	−.053	(.084)
SATQ-2	−.568	(.056)	−.234	(.081)
SATQ-1	−.723	(.058)	−.255	(.083)
Missing SATQ	−.470	(.052)	−.182	(.076)

Note: MNAR = missing not at random; SATQ = SAT quantitative score.
The estimates on the left are for the score model, while the estimates from the attendance model are on the right.

TABLE 3
Quartiles of Calculus 2 Future-Year Teacher Effects From MNAR-t (Top) Versus MAR (Left)

Quartile	1	2	3	4
1	34	11	2	1
2	11	20	14	2
3	3	14	19	11
4	0	2	12	33

Note: MAR = missing at random; MNAR = missing not at random.

By contrast, Lockwood et al. (2007), considering precision as well as ranking, only declare teacher effects as below/above average if their 90% confidence (posterior credible) intervals are strictly below/above 0. The difference between MAR and MNAR-t is not as strong using this approach (see Table 4), but some teachers still change categories under this more stringent criterion.

Following the suggestion of Molenberghs et al. (2008), we compare the fit of MNAR-t to that of MAR to see which classroom effects are most affected by the joint modeling of the attendance mechanism. The large amount of missing data in

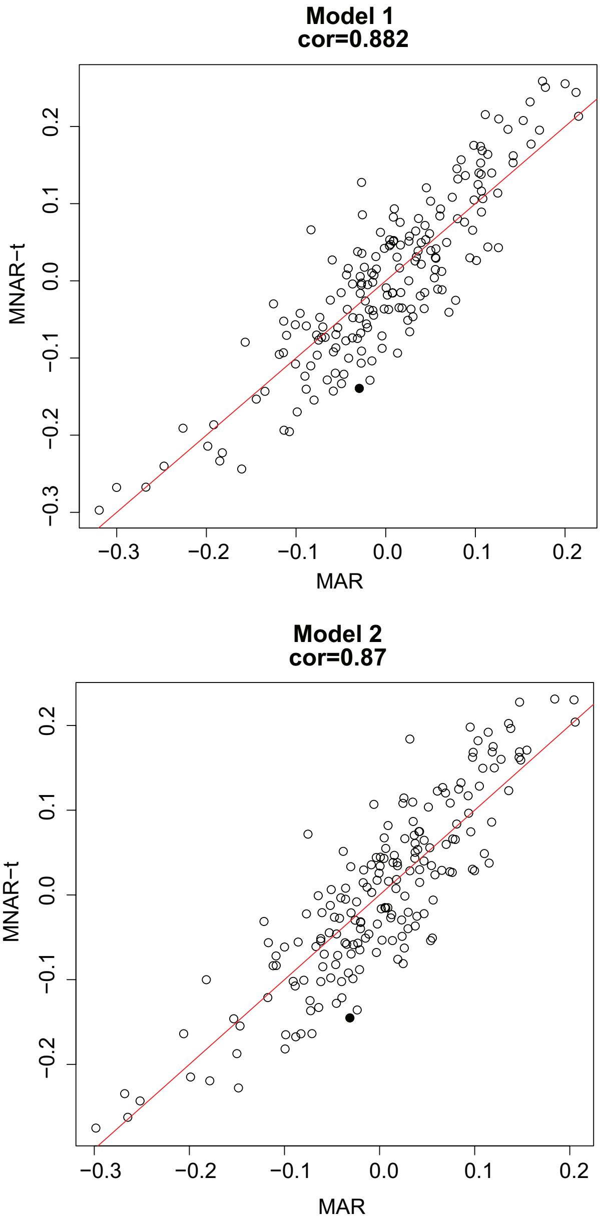

FIGURE 1. *Calculus 2 future-year effects: MAR versus MNAR-t. The solid circle represents a teacher whose value-added model score changes substantially under different assumptions for the missing data mechanism. MAR = missing at random; MNAR = missing not at random.*

TABLE 4
90% Confidence Interval Rankings for Calculus 2 Future-Year Teacher Effects From MNAR-t (Top) Versus MAR (Left)

	−	0	+
−	5	2	0
0	2	171	7
+	0	0	2

Note: MAR = missing at random; MNAR = missing not at random.

certain calculus classrooms means that the effects of those classrooms are attenuated toward zero due to the shrinkage properties of EBLUPs. This shrinkage property is normally desirable in VAMs, but in the case of potentially nonignorable dropout, we may lose information. For illustration, we examine the records of one of the teachers most greatly down-weighted by MNAR-t in Figure 1. This teacher's effect changed from -0.03 under MAR to -0.14 under MNAR-t in Model 1, and is represented by the solid circle in Figure 1. Only 20% of the students from this classroom completed Calculus 3 (most of them failed the Calculus 2 course), and those who did all received below-average grades in their respective Calculus 3 classrooms. The Calculus 2 teacher's effect on Calculus 3 in the MAR model is less than zero, but is severely shrunk because only a few observations are present. It is possible that the poor performance of this teacher's students was due entirely to student attributes that were not included in the model: motivation, major, time of course during day, and so on. However, this example illustrates how exploring the sensitivity of effects to the attendance mechanism can lead to different conclusions about teachers.

The correlation matrix for the effects of Calculus 2 teachers from Model 1 under MNAR-t appears in Figure 3. The last column of these matrices, "3 completion," yields information about the correlation of the attendance effect of the Calculus 2 teachers. A larger attendance effect means that relatively more of a teacher's students go on to complete Calculus 3. This effect is positively correlated with both the "2 on 2" effect and the "2 on 3" effect, so that the attendance effect is correlated with high grades of the teacher's students in both Calculus 2 and Calculus 3. However, the current- and future-year effects for Calculus 2 teachers are not correlated. For this data set, observing that a teacher gives above- or below-average grades yields no information about how well the students of that teacher perform in Calculus 3. Applications of VAMs to standardized test score data in the elementary school setting usually show a strong positive correlation between the current- and future-teacher effects (Karl et al., 2013b; Mariano et al., 2010).

The correlations ρ for Model 2 are nearly identical to those for Model 1 appearing in Table 3. The correlations between MAR and MNAR-t, MNAR-b, and MNAR-s, for Model 2 are 0.870, 0.968, and 0.992, respectively.

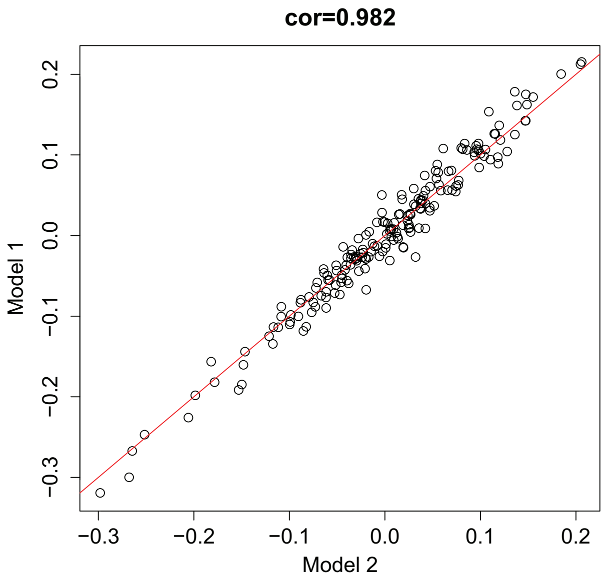

FIGURE 2. *Calculus 2 future-year effects: Model 1 MAR versus Model 2 MAR. MAR = missing at random.*

$$\mathrm{cor}(\boldsymbol{\Gamma}_1) = \begin{array}{c} \\ 2\text{ on }2 \\ 2\text{ on }3 \\ 3\text{ comp.} \end{array} \begin{array}{c} \begin{array}{ccc} 2\text{ on }2 & 2\text{ on }3 & 3\text{ completion} \end{array} \\ \begin{pmatrix} 1 & -0.028 & 0.746 \\ -0.028 & 1 & 0.596 \\ 0.746 & 0.596 & 1 \end{pmatrix} \end{array}$$

FIGURE 3. *Correlation matrix of Calculus 2 teacher effects from Model 1 under MNAR-t. "2 on 2" represents the effect of the Calculus 2 teachers on Calculus 2 grades, and "2 on 3" represents their effect on Calculus 3 grades. "3 Completion" gives the effect of Calculus 2 teachers on Calculus 3 attendance. MNAR = missing not at random.*

Furthermore, the fixed-effects parameter estimates for Model 2 under MAR were nearly identical to those obtained under MNAR-t. The estimates appear in Table 4. Figure 2 compares the teacher ratings for Models 1 and 2 under an assumption of MAR. Interestingly, the addition of significant fixed effects to the model did not have a large impact on the EBLUPs. This contrasts with the difference seen between the rankings for Model 1 (and likewise Model 2) under MAR and MNAR-t seen in Figure 1.

Figure 4 compares the student score effects from Model 2 under assumptions MAR and MNAR-b (the results are nearly identical when comparing MAR and MNAR-s). Under MNAR-b, students who attended both years of calculus saw

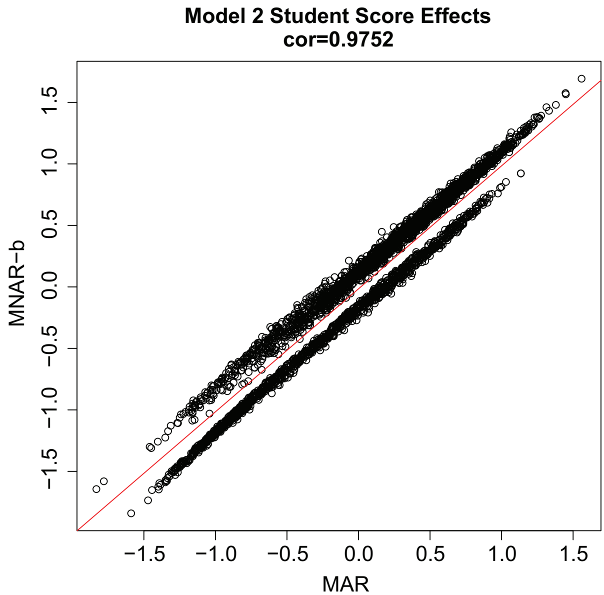

FIGURE 4. *Comparing the student score effects from MAR and MNAR-b under Model 2. MAR = missing at random; MNAR = missing not at random.*

their score effect increase under MNAR-b, while those who attended only Calculus 2 had their effects decreased. Figure 5 shows the near-perfect correlation of student score and attendance effects in Model 2 under MNAR-b. Since there is only 1 year of observations (Calculus 3) modeled by the attendance mechanism, the student attendance effects must borrow strength from the student score effects in order to be estimated. From Figure 5, it appears that these effects are perfectly correlated. This is the same result we would have obtained for the student attendance effects if we had used a shared parameter model rather than a CPM. Under the CPM, we would expect the correlation between these effects to decrease in situations where the attendance mechanism models more than a single year of observations. For the calculus example, the inclusion of student attendance effects under MNAR-b and MNAR-s requires an assumption that those effects will be perfectly correlated with the student score effects.

The sensitivity analysis illustrates the influence that assumptions about the nature of missing data may have on the resulting teacher rankings. A challenge with MNAR models is that their fit for the missing data cannot be tested empirically. The fact that the likelihood for MNAR-b is larger than that of MNAR-t indicates that MNAR-b provides a better fit for the observed data ($\mathbf{y}^o$, $\mathbf{r}$). It does not, however, indicate a better fit for the missing data $\mathbf{y}^m$. It is entirely possible that MNAR-t provides a better fit to $\mathbf{y}^m$ than MNAR-b: perhaps MNAR-b

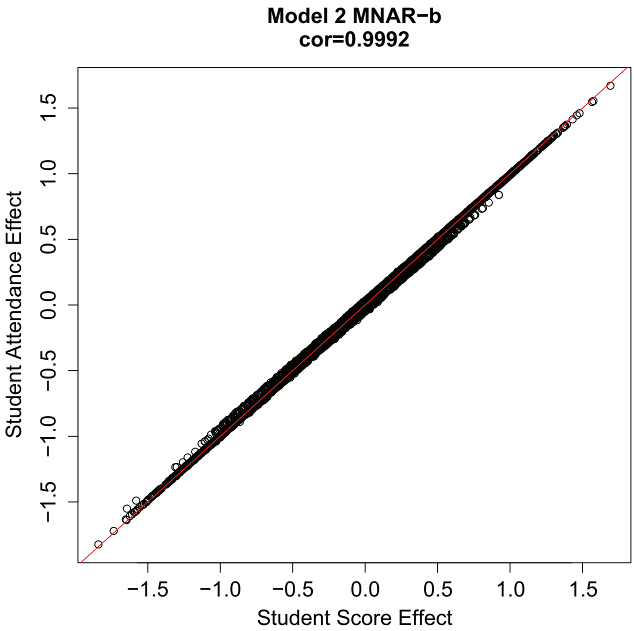

FIGURE 5. *Comparing the student score and attendance effects from MNAR-b under Model 2. MNAR = missing not at random.*

overfits $(\mathbf{y}^o, \mathbf{r})$. Without the ability to test the fit of the model to $\mathbf{y}^m$, the choice between MAR, MNAR-t, or any other relationship between the longitudinal and attendance processes requires an unverifiable assumption about the missing data process. It is interesting that the teacher effects in the score model are affected by the inclusion of teacher effects in the attendance model, but then return to their MAR values with the further inclusion of student effects in the attendance mechanism. This could represent a failure of the CI assumption for the model MNAR-t (Yuan & Little, 2009). Nevertheless, the difference in teacher effects obtained between MAR and MNAR-t demonstrates how MAR estimates may be sensitive to some MNAR modifications while robust to others.

This is a nonstandard application of a VAM; typically, these models are applied to standardized test scores from elementary and secondary students, not to university data. Furthermore, inference usually focuses on the current-year VAM effects. In this analysis, we focused on the future year effect from the GP VAM rather than the current-year effects. Ballou et al. (2004), Lockwood et al. (2007), and Mariano et al. (2010) note that the effects from the first year included in the study are susceptible to bias due to nonrandom classroom assignment and capture the cumulative effects of prior teachers on those students.

TABLE 5
Estimates From MNAR-t for Elementary School Data

	$f(\mathbf{y})$		$f(\mathbf{r})$	
μ_4^y and μ_4^r	24.303	(.167)	1.236	(.097)
μ_5^y and μ_5^r	25.289	(.167)	1.225	(.094)
μ_6^y and μ_6^r	26.315	(.172)	1.320	(.099)
σ_4^2	1.489	(.079)	—	—
σ_5^2	1.028	(.064)	—	—
σ_6^2	1.633	(.081)	—	—
Γ_{stu}	3.899	(.131)	—	—
Female	—	—	—	—
Male	0.039	(.082)	0.062	(.050)
Asian	1.500	(.226)	0.027	(.124)
Black	—	—	—	—
Hispanic	0.101	(.169)	0.346	(.092)
Native American	0.104	(.347)	−0.190	(.173)
White	1.185	(.158)	0.356	(.086)

Note: MNAR = missing not at random.
The estimates on the left are for the score model, while the estimates from the attendance model are on the right.

As with any observational data set, caution must be exercised when interpreting the results. Students were not randomly assigned to teachers, so effects ascribed to teachers may in fact be due to other factors. If students from majors that did not require Calculus 3 tended to take calculus from certain instructors, then the attendance effects of those instructors would reflect the majors of their students rather than an impact of the teacher on taking Calculus 3. We did not find evidence of clustering by major in the data set, but it is possible that time of day or other confounding factors may contribute to the estimated teacher effects.

4. Elementary School Application

We fit a different missing data mechanism to data from a large urban elementary school district. The data set tracks a cohort of 2,834 students from Grades 4 through 6, recording their score on a standardized math test each year. The data set contains 102, 104, and 98 fourth-, fifth-, and sixth-grade teachers, respectively. Fixed effects representing the mean response in each year, race/ethnicity, and gender are included in both the score and the attendance mechanisms. In the elementary setting, students typically have no choice about whether to progress to the next grade. In this setting, we would not expect grade g teachers to have an effect on whether their students take the test in grade $(g+1)$, but they might have an effect on whether their students take the test in grade g. We therefore fit a different model for the attendance process than for the university data. In this

$\mathbf{\Gamma}_4$:

$$\begin{pmatrix} 0.648 & 0.349 & 0.332 & 0.120 \\ 0.349 & 0.225 & 0.219 & 0.099 \\ 0.332 & 0.219 & 0.238 & 0.077 \\ 0.120 & 0.099 & 0.077 & 0.099 \end{pmatrix} \begin{pmatrix} 1.000 & 0.914 & 0.846 & 0.474 \\ 0.914 & 1.000 & 0.947 & 0.660 \\ 0.846 & 0.947 & 1.000 & 0.498 \\ 0.474 & 0.660 & 0.498 & 1.000 \end{pmatrix}$$

$\mathbf{\Gamma}_5$:

$$\begin{pmatrix} 0.412 & 0.165 & 0.025 \\ 0.165 & 0.084 & 0.012 \\ 0.025 & 0.012 & 0.060 \end{pmatrix} \begin{pmatrix} 1.000 & 0.889 & 0.157 \\ 0.889 & 1.000 & 0.165 \\ 0.157 & 0.165 & 1.000 \end{pmatrix}$$

$\mathbf{\Gamma}_6$:

$$\begin{pmatrix} 0.441 & 0.111 \\ 0.111 & 0.112 \end{pmatrix} \begin{pmatrix} 1.000 & 0.500 \\ 0.500 & 1.000 \end{pmatrix}$$

FIGURE 6. *Estimated blocks of the **G** matrix from MNAR-t. The covariance matrix is on the left and the correlation matrix is on the right. Within each matrix, the current-year score effects appear in the leftmost column, followed by future year score effects, and then by the current-year attendance effect. MNAR = missing not at random.*

model, $\Lambda_{g[j]}$ represents the effect that the *j*th grade *g* teacher has on the probability of his or her students being measured in the same year *g*. A total of 421 of the 6,657 student observations with recorded teacher links are missing a test score.

Despite finding moderate correlations between the teacher effects in the score and attendance models (see Table 5 and Figure 6 for the parameter estimates), the estimates of teacher effects on scores are practically identical under each model adopted to explore the missing data mechanism. The correlations between teacher effects under MAR and MNAR-t are all greater than 0.992; the plots are not displayed here because they are essentially straight lines. We also fit the model used in Section 3, exploring possible teacher effects on attendance in the following year, and likewise find that the model adopted for the missing data make little difference to the estimates of teacher effects on scores. This could be related to the fact that only around 6% of the observations are missing. By contrast, around 40% of the observations in the calculus example were missing.

In this data set from an elementary school district, the estimates of teacher effects on scores are insensitive to the choice of attendance mechanism (from those that were presented), though this does not imply that the missing data mechanism is ignorable. This insensitivity may also be a function of the relatively small proportion of missing data in this example. Graham (2009) observes that all missing data are on a continuum between MAR and MNAR: We should focus on whether or not the likely violations of MAR matter to any practical extent. Even in such situations when the teacher effects do not show sensitivity to the choice of several different MNAR models, this class of correlated random effects models may still be useful for searching for abnormal features of the data set. For example, unlike in the university setting, we might not expect to see a strong relationship between current-year teacher effects and next-year attendance effects. Yet, some teachers might appear to be outliers in bivariate plots of these effects, giving information about unusual cases in the data. As always, these potential outliers may be due to confounding factors, but they may indicate teachers with an unusual pattern.

5. Summary

We have developed a correlated random effects model to explore the sensitivity of teacher rankings from the GP VAM (Mariano et al., 2010) to assumptions about the missing data process. In an application to calculus grades from a large university, the MAR teacher effects matched those obtained from two MNAR models that allowed the attendance process to depend on random student effects. The effects were robust even in the presence of significant correlation between random effects in the score and attendance models. If a given joint model is assumed to be correct, then correlation between the longitudinal and missingness processes indicates that the missing data are nonignorable. The finding highlights the point by Graham (2009) that the focus of a sensitivity analysis should not be on whether the MAR assumption has been violated, but rather on whether the violation is large enough to have practical implications.

The joint model MNAR-t, which allows for MNAR data under the specified attendance mechanism with included teacher effects, produces a different ranking and classification of the Calculus 2 teacher future effects than the MAR GP model (the current-year effects were unaffected). By contrast, MNAR-t produced roughly the same teacher rankings as the MAR model for the elementary school example. Likewise, McCaffrey and Lockwood (2011) did not find an appreciable difference in the results of their MNAR and MAR models while analyzing data from elementary school standardized scores, attributing the missingness to student characteristics. Three important differences between the calculus and the elementary school examples are the lack of standardization in the calculus grade assignments, the larger percentage of missing data in the calculus example, and the greater potential for the calculus attendance trajectories of students to vary by teacher, due to the greater choice college students have in selecting future courses. In addition, the calculus rankings would have likely benefited from the inclusion of additional covariates such as the student major and the time of day that the course is offered. These factors may help explain the more profound changes to calculus teacher rankings resulting from the joint model MNAR-t.

In an application to elementary school data, none of the presented MNAR models produce a large number of significantly different teacher effects from those obtained under MAR. We would expect that in many elementary data settings, the teachers would have little effect on their students' attendance at the test. However, the missing data models proposed in this article could be used to identify unusual patterns in the data if such occurred. In secondary school data, one might expect to see an effect of grade g teachers on grade $(g+1)$ class taking, particularly with elective classes. For example, if the high schools require only 2 years of math, a sophomore math teacher may have an effect on his or her students' decisions to take advanced math classes. Thus, we would expect that the missing data models used for the calculus data in this article would also be useful at the secondary level.

VAMs are typically fit on observational data, not on data from a designed experiment. It is therefore always a possibility that the effects on student test scores that are ascribed to teachers are actually due to an unmeasured attribute of students who are assigned to that teacher. The same is true for the attendance models proposed in this article. In the university setting, a teacher may have a low fraction of students proceed to Calculus 3 if that teacher's students are in a discipline that does not require Calculus 3. At the elementary school level, a teacher may be assigned a class with a large number of students who are exempt from the testing requirement, in which case the data are missing because of student rather than teacher characteristics. Thus, effects estimated for individual teachers must be interpreted carefully and other potential confounding factors need to be considered.

The methodology of this article has been developed in the educational setting, but it applies in many other arenas as well. For example, longitudinal studies of medical interventions often have missing data and the patients may be treated by several medical practitioners or hospitals. The methods of this article can be used to evaluate effects of missing data in this context.

Acknowledgment

We would like to thank the three anonymous referees for their careful review and helpful suggestions that led to an improved article.

Authors' Note

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the National Science Foundation, Arizona State University, Adsurgo LLC, or Westat.

Declaration of Conflicting Interests

The author(s) declared no conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was partially supported by the National Science Foundation under grant DRL-0909630, and by Arizona State University through a Dissertation Fellowship.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, *25*, 95–135.

Allison, P. D. (2008). Convergence failures in logistic regression (SAS Global Forum 2008, no. 360-2008). Retrieved from www2.sas.com/proceedings/forum2008/360-2008.pdf

Amrein-Beardsley, A. (2008). Methodological concerns about the education value-added assessment system. *Educational Researcher*, *37*, 65–75.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L. , . . . Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers* (Economic Policy Institute Briefing Paper, 278). Retrieved from www.epi.org/publication/bp278

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

Betebenner, D. W. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practices*, *28*, 42–51.

Braun, H. (2005). Value-added modeling: What does due dilligence require? In R. Lissitz (Ed.), *Value added models in education: Theory and applications* (pp. 19–38). Maple Grove, MN: JAM Press.

Braun, H. I., Chudowsky, N., & Koenig, J. (2010), *Getting value out of value-added*. Washington, DC: National Academies Press.

Briggs, D., & Domingue, B. (2011). Due diligence and the evaluation of teachers: A review of the value-added analysis underlying the effectiveness rankings of Los Angeles unified school district teachers by the *Los Angeles Times. National Education Policy Center*. Retrieved from www.nepc.colorado.edu/publication/due–diligence

Broatch, J., & Lohr, S. (2012). Multidimensional assessment of value added by teachers to real-world outcomes. *Journal of Educational and Behavioral Statistics*, *37*, 256–277.

Fernandez, M. (2012). El Paso schools confront scandal of students who disappeared at test time. Retrieved from http://www.nytimes.com/2012/10/14/education/el-paso-rattled-by-scandal-of-disappeared-students.html

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis*. Hoboken, NJ: John Wiley.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549–576.

Hanushek, E. A. (1971). Teacher characteristics and gains in student achievement: Estimation using micro data. *American Economic Review*, *60*, 280–288.

Harris, D. N. (2011). Value-added measures and the future of educational accountability. *Science*, *333*, 826–827.

Harris, D. N., & McCaffrey, D. F. (2010). Value-added: Assessing teachers' contributions to student achievement. In M. M. Kennedy (Ed.), *Teacher assessment and the quest for teacher quality* (pp. 251–282). San Francisco, CA: Jossey-Bass.

Karl, A. T., Yang, Y., & Lohr, S. (2012). GPvam: Maximum likelihood estimation of multiple membership mixed models used in value-added modeling. Retrieved from http://cran.r-project.org/web/packages/GPvam/index.html, R package version 2.0-0

Karl, A. T., Yang, Y., & Lohr, S. L. (2013a). Computation of maximum likelihood estimates for multiresponse generalized linear mixed models with non-nested, correlated random effects. Submitted.

Karl, A. T., Yang, Y., & Lohr, S. L. (2013b). Efficient maximum likelihood estimation of multiple membership linear mixed models, with an application to educational value-added assessments. *Computational Statistics and Data Analysis*, *59*, 13–27.

Koretz, D. M. (2008). *Measuring up: What educational testing really tells us*. Cambridge, MA: Harvard University Press.

Lin, H., Liu, D., & Zhou, X.-H. (2009). A correlated random-effects model for normal longitudinal data with nonignorable missingness. *Statistics in Medicine*, *29*, 236–247.

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

Lockwood, J., McCaffrey, D., Mariano, L., & Setodji, C. (2007). Bayesian methods for scalable multivariate value-added assessment. *Journal of Educational and Behavioral Statistics*, *32*, 125–150.

Lohr, S. (2012). The value Deming's ideas can add to educational evaluation. *Statistics, Politics, and Policy*, *3*, article 1.

Mariano, L. T., McCaffrey, D. F., & Lockwood, J. (2010). A model for teacher effects from longitudinal data without assuming vertical scaling. *Journal of Educational and Behavioral Statistics*, *35*, 253–279.

McCaffrey, D. F., & Lockwood, J. R. (2011). Missing data in value-added modeling of teacher effects. *Annals of Applied Statistics*, *5*, 773–797.

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

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability*. Pittsburgh, PA: The RAND Corporation.

McCaffrey, D. F., Lockwood, J. R., Mariano, L. T., & Setodji, C. (2005). Challenges for value- added assessment of teacher effects. In R. Lissitz (Ed.), *Value added models in education: Theory and applications* (pp. 111–144). Maple Grove, MN: JAM Press.

McCulloch, C. E. (1994). Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association*, *89*, 330–335.

Molenberghs, G., & Kenward, M. G. (2007). *Missing data in clinical studies*. West Sussex, England: John Wiley.

Molenberghs, G., Kenward, M. G., Verbeke, G., Beunckens, C., & Sotto, C. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society, Series B*, *70*, 371–388.

The National Academies. (2009). Letter report to the U.S. Department of Education on the Race to the Top fund. Retrieved from www.nap.edu/catalog/12780.html

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

R Development Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rhoads, C. H. (2012). Problems with test of the missingness mechanism in quantitative policy studies. *Statistics*, *Politics, and Policy*, *3*, Article 6.

Rizopoulos, D., Verbeke, G., & Lesaffre, E. (2009). Fully exponential Laplace approximations for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society, Series B*, *71*, 637–654.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics*, *125*, 175–214.

Rowan, B., Correnti, R., & Miller, R. J. (2002). What large-scale survey research tells us about teacher effects on student achievement: Insights from the prospects study of elementary schools. *Teachers College Record*, *104*, 1525–1567.

Ryan, R. M., & Weinstein, N. (2009). Undermining quality teaching and learning: A self-determination theory perspective on high-stakes testing. *Theory and Research in Education*, *7*, 224–233.

Sanders, W., Saxton, A., & Horn, S. (1997). The Tennessee value-added assessment system: A quantitative outcomes-based approach to educational assessment. In J. Millman (Ed.), *Grading teachers, grading schools. Is student achievement a valid evaluation measure?* (pp. 137–162). Thousand Oaks, CA: Corwin Press.

Steele, B. M. (1996). A modified EM algorithm for estimation in generalized mixed models. *Biometrics*, *52*, 1295–1310.

Wright, S. P. (2004). *Advantages of a multivariate longitudinal approach to educational value-added assessment without imputation*. Presented at the 2004 National Evaluation Institute, July 8-10, 2004, Colorado Springs, CO. Retrieved from www.sas.com/resources/asset/educational–value–added–assessment.pdf

Wright, S. P., White, J. T., & Sanders, W. L. (2010). *SAS EVAAS statistical models*. Cary, NC: SAS Institute. Retrieved from www.sas.com/resources/asset/SAS-EVAAS-Statistical-Models.pdf

Wu, M. C., & Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, *44*, 175–188.

Xu, S., & Blozis, S. A. (2011). Sensitivity analysis of mixed models for incomplete longitudinal data. *Journal of Educational and Behavioral Statistics*, *36*, 237–256.

Yuan, Y., & Little, R. J. A. (2009). Mixed-effect hybrid models for longitudinal data with nonignorable dropout. *Biometrics*, *65*, 478–486.

Authors

ANDREW T. KARL is a statistician with Adsurgo LLC, 3700 Quebec St., Unit 100, Suite 258, Denver, CO 80207; email: andrew.karl@adsurgo.com. His research interests include computation and applications of linear and nonlinear mixed models, parallel statistical computing, and value-added models for teacher effects.

YAN YANG is an adjunct faculty member at Arizona State University, School of Mathematical and Statistical Sciences, Tempe, AZ 85287; email: yan.yang@asu.edu. Her areas of interest include finite mixture models, statistical computing, missing data, and statistical models for educational evaluation.

SHARON L. LOHR is a vice president at Westat, 1600 Research Boulevard, Rockville MD 20850; email: sharonlohr@westat.com. Her areas of interest include survey sampling, design of experiments, mixed models, and value-added modeling of teacher effects.

Manuscript received August 31, 2012
Revision received February 8, 2013
Accepted April 25, 2013