Cross-Classified Models in the Context of Value-Added Modeling

J. Kyle Roberts
Southern Methodist University

Douglas Bates
University of Wisconsin-Madison

1 Background and Context

In the past few years, value-added modeling (VAM) has become increasingly popular as a means for monitoring student knowledge growth and teacher accountability. Initially, VAMs focused on model development (McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Tekwe et al., 2004) and the development of software routines appropriate for estimating these models. Presently, researchers are beginning to question other aspects of VAMs, such as whether these estimates can reasonably be thought of as causal (Rubin, Stuart, & Zanutto, 2004) and the appropriateness of cross-classified models for teacher accountability (Lockwood, Doran & McCaffrey, 2003).

Work by Doran, Bates, and Phillips (under review) and work presented in the special issue of the Journal of Educational and Behavioral Statistics, volume 29(1) have made considerable strides above and beyond initial VAMs. Specifically, Doran et al. are building models which honor the psychometric properties of the instruments used in VAMs. Furthermore, their findings indicate that the VAM school effects are not invariant to the choice of an item response theory model.

If models indeed are going to be built for teachers (and schools and districts), a cross-classified model (Lockwood, Doran & McCaffrey, 2003; Raudenbush & Bryk, 2002) is preferred over simpler multilevel models. This class of statistics models the property that teacher (or school and district) effects have an additive characteristic to student continual performance, such that past teachers may affect future outcomes. The difficulty is that these types of models are complex to build and computationally demanding. Recent developments in statistical software routines have also paved the way for the development of more complex VAMs. Developments in packages like `lme4` in the R software package have made these more complex models easily estimable through advances in the sparse Cholesky decomposition (Bates, 2008). The strength with handling data in this type of way is that prior and future effects of teachers/schools/districts can be modeled as random effects and their potential value to a student's growth towards adequate yearly progress can be correctly modeled. Not honoring this model structure could inadvertently inflate (or deflate) the perceived value-added for a given teacher if students coming in to the teacher's class previously had highly effective (or less effective) teachers.

One of the first true VAMs was introduced by Sanders, Saxton & Horn (1997). In this work, the Tennessee Value-Added Assessment System was first articulated, but the authors quickly noted that a model like this had tremendous computational requirements. Although great strides have been made in the area of software development, Lockwood et al. (2007) has recently noted that there are still computational problems for solving VAMs for large datasets and hence large districts and states.

VAMs typically fall into one of two categories: models appropriate for monitoring school effects on student learning and models appropriate for monitoring teacher effects on student learning. These models typically take the form:

$$Y_{ti} = [\beta_0 + \beta_1(year_t)] + \left[\theta_{0j(i)} + \theta_{1j(i)}(year_t) + \delta_{0i} + \delta_{1i}(year_t) + \epsilon_{ti}\right] \tag{1}$$

where i indexes students and j could index either the schools or teachers. The notation $j(i)$ represents the school (or teacher) attended by student i in year t. In this model, the within-group errors are assumed IID where $\epsilon_{ti} \sim \mathcal{N}(0, \sigma^2)$ and school (or teacher) and student random effects are $\boldsymbol{\theta}_j = (\theta_{0j}, \theta_{1j}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$ and $\boldsymbol{\delta}_i = (\delta_{0i}, \delta_{1i}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega})$, respectively.

Furthermore, covariates may be added as fixed effects to Equation 1 such that:

$$Y_{ti} = [\beta_0 + \beta_1(year_t) + \beta_q x_{qti}] + \left[\theta_{0j(i)} + \theta_{1j(i)}(year_t) + \delta_{0i} + \delta_{1i}(year_t) + \epsilon_{ti}\right] \tag{2}$$

where x_{qti} represents the vector of all student-level covariates for student i at time point t. In this case, all of the covariates are modeled as fixed effects.

There is a subtle, but important, oversight in our notation in equations (1) and (2) that illustrates the computational challenges in fitting VAMs to large longitudinal (i.e. following students over time) data sets. We write $j(i)$ for the school (or teacher's class) attended by student i at time t whereas it should be written $j(t, i)$. The data sets described in section 2.3 include results from several years of each student's career during which time most students will change schools and nearly all students will change teachers. So if j is the teacher index then the expression $j(i)$ is not well-defined because student i has different teachers in different years.

McCaffrey, Lockwood, Koretz & Hamilton (2003) raised some serious issues as to the usability of the results from VAMs. Among the questions introduced in their research are the problems concerning omitted variables and missing data, problems concerning the type and nature of the instruments used to measure student achievement, problems with modeling measurement error, and problems with using achievement measures as proxies for estimating teacher effectiveness. As part of their research, McCaffrey et al. (2003) made seven recommendations for future research:

1. Develop databases that can support VAM estimation of teacher effects.
2. Develop computational tools for fitting VAM
3. Link VAM teacher-effect estimates to alternative measures of teacher effectiveness
4. Empirically evaluate the potential sources of errors we have identified
5. Estimate the prevalence of factors that contribute to the sensitivity of teacher-effect estimates
6. Incorporate decision theory into VAM
7. Use research and auxiliary data to inform modeling choices (pp. 114-119)

Several researchers have already begun working to contribute to answering some of these calls for research. For example, Doran, Bates, and Phillips (under review) have turned their attention to investigating the consequences of psychometric decisions on VAMs for school effects (Recommendation 4). In the current paper, we turn our attention specifically toward Recommendation 2 and the development of the `lme4` package in R.

2 Purpose

Cross-classified models are a special case of mixed-effects models in which the typical assumption of complete nesting in hierarchical linear modeling is not met. For data to be considered purely clustered, clusters of elements from one level must all belong to single elements of a higher level. For example, in the following table, students from five separate high schools are completely nested inside each of the five high schools.

	School 1	School 2	School 3	School 4	School 5
Student	1,2,3	4,5	6,7,8	9,10	11,12,13,14

In the above example, each student is nested inside one and only one school. This data may be modeled in a mixed-effects model as:

$$Y_{ij} = \gamma_{00} + u_{0j} + e_{ij} \tag{3}$$

where γ_{00} represents the average student achievement across all j schools with student-level residuals e_{ij} and school-level random effects u_{0j}. This type of conventional modeling is sometimes referred to as hierarchical linear modeling (Raudenbush & Bryk, 2002), multilevel modeling (Hox, 2002), and mixed-effects modeling (Pinheiro & Bates, 2000).

However, consider an effect where student achievement scores are modeled as a product of both which middle school they attend and which high school they attend. In an ideal setting, students from each middle school would matriculate into one and only one high school. In this case, the data might be represented as:

	Middle School					
High School	MS 1	MS 2	MS 3	MS 4	MS 5	MS 6
HS 1	S1, S2	S3, S4, S5				
HS 2			S6, S7	S8, S9, S10		
HS 3					S11, S12, S13	S14, S15

In this example, each student is nested inside one and only one middle school and these middle schools are nested inside three high schools. This data may be modeled in a mixed-effects model as:

$$Y_{ijk} = \gamma_{000} + u_{00k} + u_{0jk} + e_{ijk} \tag{4}$$

where γ_{000} represents the average student achievement across all j middle schools and k high schools with student-level residuals e_{ijk} and middle school-level random effects u_{0jk} and high school-level random effects u_{00k}.

Although the above representation is an ideal setting, it is more the convention that students are actually cross-classified by both middle school and high school such that their data structure resemble:

	Middle School					
High School	MS 1	MS 2	MS 3	MS 4	MS 5	MS 6
HS 1	S1, S2	S3	S4		S5	
HS 2		S6, S7		S8, S9, S10		
HS 3	S11		S12		S13	S14, S15

In this cross-classified example, student scores are modeled as:

$$Y_{i(j_1,j_2)} = \gamma_{000} + u_{00j_1} + u_{00j_2} + u_{00j_1xj_2} + e_{i(j_1,j_2)} \tag{5}$$

where γ_{000} represents the average student achievement for student i having attended middle school j_1 and high school j_2 with random effects u_{00j_1} for the middle schools, u_{00j_2} for the high schools, a random interaction effect $u_{00j_1xj_2}$ for belonging to both middle school j_1 and high school j_2 and student-level residuals $e_{i(j_1,j_2)}$. This parameterization follows the Rasbash and Browne (2001) notation.

This type of cross-classified modeling has strong applications in monitoring teacher and school effects for student achievement scores across time. In the case where the object of modeling is teacher effectiveness, the cross-classified model is the only convention for modeling time series data unless the measurement occasions are limited to just one year. Consider just the simple case where two years worth of student data (7th and 8th grades) are modeled in a cross-classified setting. In this case, our model would take on the form:

$$Y_{ti(j_1,j_2)} = \gamma_{0000} + u_{000j_1} + u_{000j_2} + u_{000j_1xj_2} + u_{00ij_1} + u_{00ij_2} + u_{00ij_1xj_2}e_{ti(j_1,j_2)}. \tag{6}$$

As can be seen from the above equation, these models become exponentially complex the more years worth of data and the more grades we add to a model. Indeed, Lockwood et al. (2007) note that these models can become so increasingly complex that more efficient algorithms need to be developed in order to conduct further research into cross-classified models. They also propose a notation slightly different from the Rasbash and Browne (2001) notation where teacher effects are modeled as:

$$Y_{ti} = \mu_t + \beta_t^{'}x_{it} + \sum_{t^* \leq t} \alpha_{tt^*}\phi_{it^*}^{'}\theta_{t^*} + \epsilon_{it} \tag{7}$$

where μ_t is the overall mean for each year, x_{it} is the covariate vector of both time invariant and time varying variables, θ_t are the teacher effects for each year, ϕ_{it} are the student effects vectors, and ϵ_{ti} is the residual score for student i at time *t*.

3 Setting, Population, and Intervention

This data is simulated, so there is no immediate setting, population or intervention.

4 Research Design

Suppose we use random effects for students, teachers and schools in a model for data with a total of n_1 students, n_2 teachers and n_3 schools. If we allow for two random effects for students (the random effect for the intercept and for the slope with respect to time — δ_{0i} and δ_{1i} in equations (1) and (2)) then the model matrix $\boldsymbol{Z}_1$ for the student random effects will be $n \times 2n_1$. The random-effects model matrices $\boldsymbol{Z}_2$ for teachers and $\boldsymbol{Z}_3$ for schools would have dimension $n \times 2n_2$ and $n \times 2n_3$, respectively. Let $\boldsymbol{b}$, a vector of length $q = 2(n_1 + n_2 + n_3)$, be the vector of all random effects (students, teachers and schools) and $\boldsymbol{Z} = [\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{Z}_3]$ be the corresponding model matrix.

The value-added model could then be written

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zb} + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n),\ \boldsymbol{b} \sim \mathcal{N}\left(\boldsymbol{0}, \Sigma(\boldsymbol{\theta})\right) \tag{8}$$

where $\boldsymbol{\theta}$ is the vector of variance components to be estimated. For the model we are describing $\boldsymbol{\theta}$ would have a total of nine elements; two variances and one covariance for each of the student, teacher and school classifications.

Because the matrix $\boldsymbol{Z}$ can be huge (perhaps millions of rows and hundreds of thousands of columns) this representation would be interesting but completely impractical if it were not for the fact that $\boldsymbol{Z}$ is very, very sparse. In the example we are describing each row is zero in all but six of the columns (two student random effects, two teacher random effects and two school random effects). Such matrices can be stored and manipulated using sparse matrix techniques (Davis, 2006), even for large databases.

One of the shining successes of sparse matrix theory and software is the ability to calculate the Cholesky decomposition of large, sparse, symmetric, positive-definite matrices. Fortuitously, the calculations needed to evaluate the log-likelihood for the parameters in a model like (8) can be reduced to exactly such a calculation. Without going into details, the broad outlines are to express the variance-covariance of the random effects, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, in terms of a relative covariance factor, $\boldsymbol{\Lambda}(\boldsymbol{\theta})$, defined so that

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \sigma^2 \Lambda(\boldsymbol{\theta})\boldsymbol{\Lambda}'(\boldsymbol{\theta}) \tag{9}$$

and optimize the profiled deviance, defined as

$$-2\tilde{\ell}(\boldsymbol{\theta}) = \log(|\boldsymbol{L}(\boldsymbol{\theta})|^2) + n\left[1 + \frac{2\pi r^2(\boldsymbol{\theta})}{n}\right] \tag{10}$$

where $\boldsymbol{L}(\boldsymbol{\theta})$ is the sparse, lower-triangular Cholesky factor satisfying

$$\boldsymbol{L}(\boldsymbol{\theta})\boldsymbol{L}'(\boldsymbol{\theta}) = \boldsymbol{\Lambda}(\boldsymbol{\theta})'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Lambda}(\boldsymbol{\theta}) + \boldsymbol{I}. \tag{11}$$

In (10) $|\boldsymbol{L}(\boldsymbol{\theta})|$ denotes the determinant of $\boldsymbol{L}$, which is simply the product of its diagonal elements, and $r^2(\boldsymbol{\theta})$ is the minimum penalized residual sum of squares, which can be written as

$$r^2(\boldsymbol{\theta}) = \min_{\boldsymbol{u},\boldsymbol{\beta}} \left\| \begin{bmatrix} \boldsymbol{y} \\ 0 \end{bmatrix} - \begin{bmatrix} \boldsymbol{Z}\boldsymbol{\Lambda}(\boldsymbol{\theta}) & \boldsymbol{X} \\ \boldsymbol{I} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\beta} \end{bmatrix} \right\|^2 . \tag{12}$$

Calculating the Cholesky factor, $\boldsymbol{L}(\boldsymbol{\theta})$, is the crucial step in solving the penalized least squares problem (12).

This paper will show the consequences of neglecting this cross-classified data structure in VAMs when data are modeled as fully nested and the cross-classification is ignored.

5 Findings

In one of the first evaluations of different VAMs, Tekwe et al. (2004) setup their process of critical appraisal by considering four separate types of VAMs: the simple change score fixed effects model, the simple unadjusted hierarchical linear mixed model, the hierarchical linear mixed model adjusted for student- and school-level covariates, and the layered mixed effects model. The results of their study led them to prefer the simpler change score fixed effects model over more complicated models, although it should be noted that their approach did not actually provide any simulation-type estimates and was limited to observations of the models.

In the final phase of analysis, we compare our cross-classified model to the other types of models described above. In this phase, show that previous VAMs incorrectly model variance components and as a result “can” lead to erroneous results.

Références

Bates, D. M. (2008). Computational methods for mixed models. Technical report, University of Wisconsin - Madison, http://cran.cnr.berkeley.edu/.

Davis, T. A. (2006). *Direct methods for sparse linear systems*. SIAM, Philadelphia, PA.

Doran, H., Bates, D. M., and Phillips, G. (under review). The consequences of various psychometric decisions on school effects estimated from value-added models.

Goldstein, H. (1995). Hierarchical data modeling in the social sciences. *Journal of Educational and Behavioral Statistics*, 20:201–204.

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. Erlbaum, Mahwah, NJ.

Kreft, I. and de Leeuw, J. (1999). *Introducing multilevel modeling*. Sage, Thousand Oaks, CA.

Lockwood, J., Doran, H., and McCaffrey, D. F. (2003). Using r for estimating longitudinal student achievement models. *The Newsletter of the R Project*, 3(3):17–23.

Lockwood, J., McCaffrey, D. F., Mariano, L. T., and Setodji, C. (2007). Bayesian methods for scalabile multivariate value-added assessment. *Journal of Educational and Behavioral Statistics*, 32(2):125–150.

McCaffrey, D. F., Lockwood, J., Koretz, D., Louis, T., and Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, 29(1):67–101.

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

Pinheiro, J. and Bates, D. M. (2000). *Mixed-effects models in S and S-Plus*. Springer, New York.

Rasbash, J. and Browne, W. J. (2001). *Multilevel modeling of health statistics*, chapter Modeling non-hierarchical structures, pages 93–105. John Wiley and Sons, Chichester.

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

Rubin, D., Stuart, E. A., and Zanutto, E. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics*, 29(1):103–116.

Sanders, W. L., Saxton, A., and Horn, S. (1997). *Grading teachers, grading schools: Is student achievement a valid evaluation measure?*, chapter The Tennessee value-added assessment system: A quantitative, outcomes-based approach to educational assessment. Corwin Press, Thousand Oaks, CA.

Sarkar, D. (2008). *Lattice: Multivariate Data Visualization with R*. Springer, New York, NY.

Snijders, T. and Bosker, R. (1999). *Multilevel analysis*. Sage, Thousand Oaks, CA.

TEA (2009). Growth model pilot application for adequate yearly progress determinations under the no child left behind act. Revised proposal submitted to the U.S. Department of Education.

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