Journal of Educational and Behavioral Statistics
Summer 2005, Vol. 30, No. 2, pp. 109–139

Prediction in Multilevel Models

David Afshartous
University of Miami

Jan de Leeuw
University of California, Los Angeles

Multilevel modeling is an increasingly popular technique for analyzing hierarchical data. This article addresses the problem of predicting a future observable y_{*j} *in the* j*th group of a hierarchical data set. Three prediction rules are considered and several analytical results on the relative performance of these prediction rules are demonstrated. In addition, the prediction rules are assessed by means of a Monte Carlo study that extensively covers both the sample size and parameter space. Specifically, the sample size space concerns the various combinations of Level 1 (individual) and Level 2 (group) sample sizes, while the parameter space concerns different intraclass correlation values. The three prediction rules employ OLS, prior, and multilevel estimators for the Level 1 coefficients* β_j*. The multilevel prediction rule performs the best across all design conditions, and the prior prediction rule degrades as the number of groups,* J*, increases. Finally, this article investigates the robustness of the multilevel prediction rule to misspecifications of the Level 2 model.*

Keywords: *Monte Carlo, multilevel model, prediction*

1. Introduction

Prediction in multilevel models is considered in terms of forecasting unobserved (yet observable) units at the individual level. Consider the school example. After carrying out a Multilevel model analysis on some data, suppose we want to know the outcome (y) for a student not in the data set. Formally, let y_{*j} be the unknown outcome measure, say, mathematics score, for an unsampled student in the jth school. The basic problem is to predict y_{*j}. We present three main approaches to the prediction of y_{*j} and examine their performance through a simulation study that extensively covers both the sample size and parameter space. In addition, we compare these results with the corresponding results for estimation.

Although there exists an extensive and growing literature on estimation issues in multilevel models (de Leeuw & Kreft, 2002; Raudenbush & Bryk, 2002), the same cannot be said with respect to prediction. Exceptions include Rubin's (1980)

This research was supported by a grant from the National Institute for Statistical Sciences. We greatly appreciate the comments from the editor and anonymous referee that have substantially improved the quality of the article.

Law School Validity Studies article where a multilevel model without group level covariates is used to predict first-year GPA based on LSAT score; he found that predictions were improved by what he termed empirical Bayes predictors. Gray, Goldstein, and Thomas (2001) consider the problem of predicting future "value-added" performance across groups from past trends. The main result is that such prediction is unreliable.[1] However, there does not exist a full treatment of the multilevel prediction problem. Multilevel prediction is an important problem given the popularity of multilevel models in a variety of fields and the usefulness of being able to forecast future observations.

In section 1.1 we review the multilevel model, and in section 1.2 we discuss estimation in multilevel models. In section 2 we present three approaches to prediction in multilevel models, and in section 3 we describe the simulation study design with which we assess these three methods. Results and discussion are in section 4, and a brief summary is in section 5.

1.1. The Multilevel Model

Multilevel modeling is a statistical technique designed to facilitate inferences from hierarchical data. Other names such as hierarchical linear modeling, random coefficient modeling, or empirical Bayes estimation, are often employed, usually as a function of one's research discipline. Nevertheless, the basic framework is the same in each case: a given data point, y_{ij}, represents the ith observation in the jth group, for example, the ith student in the jth school for educational data; we may have J groups, where the jth group contains n_j observations. Although several levels of data may be considered, this discussion is restricted to the simple case of primary units grouped within secondary units, and we periodically refer to the applied example of students (Level 1) grouped within schools (Level 2). Within each group, we have the following Level 1 model equation:

$$\mathbf{Y}_j = \mathbf{X}_j\beta_j + \mathbf{r}_j. \tag{1}$$

Each $\mathbf{X}_j$ has dimensions $n_j \times p$, and $\mathbf{r}_j \sim N(0, \sigma^2\Psi_j)$, with Ψ_j usually taken as $\mathbf{I}_{nj}$. To be sure, these J regression equations may be estimated separately, thereby ignoring the structure in the data. A common problem with this approach, however, is that some of the groups do not contain sufficient data to produce stable estimates. In multilevel modeling, this problem is remedied by modeling some or all of the Level 1 coefficients, β_j, as random variables.[2] They may also be functions of Level 2 (school) variables:

$$\beta_j = \mathbf{W}_j\gamma + \mathbf{u}_j. \tag{2}$$

Each $\mathbf{W}_j$ has dimension $p \times q$ and is a matrix of background variables on the jth group and $\mathbf{u}_j \sim N(0, \tau)$. Clearly, because τ is not necessarily diagonal, the elements of the random vector β_j are not independent. For instance, there might exist a covariance between the slope and intercept for each regression equation. Equation 2 may

be viewed as a prior for the distribution of the Level 1 β_j, modeled as varying around a conditional grand mean $\mathbf{W}_j\gamma$ with a common variance τ, thereby expressing a judgment of similarity with respect to the groups.[3] For instance, in the school example, this expresses the reasonable judgment that schools, although unique in many ways, have certain common characteristics that may be accounted for in the modeling process. Furthermore, the separate equations for Level 1 and Level 2 data readily models/displays the relationship between variables from different levels of the data, where the magnitude of the elements of γ measure the strength of these cross-level interactions. Specifically, the group Level 2 variables may either increase or decrease the individual Level 1 coefficients. For the school example, these phenomena would be classified as "school effects."

Combining equations yields the single equation model:

$$\mathbf{Y}_j = \mathbf{X}_j\mathbf{W}_{j\gamma} + \mathbf{X}_j\mathbf{u}_j + \mathbf{r}_j, \tag{3}$$

which may be viewed as a special case of the mixed linear model, with fixed effects γ and random effects $\mathbf{u}_j$.[4] Researchers more interested in the fixed effects γ rather than the Level 1 coefficients β_j often prefer this formulation. Marginally, $\mathbf{y}_j$ has expected value $\mathbf{X}_j\mathbf{W}_j\gamma$ and dispersion $\mathbf{V}_j = \mathbf{X}_j\tau\mathbf{X}_j' + \sigma^2\mathbf{I}$. Observations in the same group have correlated disturbances, and this correlation will be larger if their predictor profiles are more alike in the metric τ (de Leeuw & Kreft, 1995). Thus, the full log-likelihood for the jth unit is

$$L_j(\sigma^2, \tau, \gamma) = -\frac{n_j}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{V}_j| - \frac{1}{2}\mathbf{d}'_j\,\mathbf{V}_j^{-1}\mathbf{d}_j, \tag{4}$$

where $\mathbf{d}_j = \mathbf{Y}_j - \mathbf{X}_j\mathbf{W}_j\gamma$. Because the J units are independent, we write the log-likelihood for the entire model as a sum of unit log-likelihoods, that is,

$$L(\sigma^2, \tau, \gamma) = \sum_{j=1}^{J} L_j(\sigma^2, \tau, \gamma). \tag{5}$$

By appropriately stacking the data for each of the J Level 2 units, we may write the model for the entire data without subscripts. Thus, we have:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{r}, \tag{6}$$

with r normally distributed with mean 0 and dispersion Ψ where

$$\mathbf{Y} = (\mathbf{Y}_1', \mathbf{Y}_2', \ldots, \mathbf{Y}_J')',$$

$$\beta = (\beta_1', \beta_2', \ldots, \beta_J')',$$

$$\mathbf{r} = (\mathbf{r}_1', \mathbf{r}_1', \ldots, \mathbf{r}_J')',$$

and

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 & 0 & \dots & 0 \\ 0 & \mathbf{X}_2 & \dots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \dots & \mathbf{X}_J \end{pmatrix} \quad \Psi = \begin{pmatrix} \Psi_1 & 0 & \dots & 0 \\ 0 & \Psi_2 & \dots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \dots & \Psi_J \end{pmatrix}.$$

We may also write the Level 2 equation in no-subscript form through similar stacking manipulations:

$$\beta = \mathbf{W}\gamma + \mathbf{u}, \tag{7}$$

where u is normally distributed with mean 0 and covariance matrix T where

$$\mathbf{W} = (\mathbf{W}_1', \mathbf{W}_2', \dots, \mathbf{W}_J')',$$

$$\mathbf{u} = (\mathbf{u}_1', \mathbf{u}_2', \dots, \mathbf{u}_J')',$$

$$\mathbf{T} = \begin{pmatrix} \tau & 0 & \dots & 0 \\ 0 & \tau & \dots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \dots & \tau \end{pmatrix}.$$

Combining equations, the entire model may be written as:

$$\mathbf{Y} = \mathbf{XW}\gamma + \mathbf{Xu} + \mathbf{r}, \tag{8}$$

where we note that $E(\mathbf{y}) = \mathbf{XW}\gamma$ and $\mathrm{Var}(\mathbf{y}) = \mathbf{XTX}' + \Psi$.

1.2. Estimation

Given that the multilevel model may be viewed from a variety of perspectives (e.g., separate equation model versus combined equation model), so can the approaches to estimation. Raudenbush and Bryk (2002) discuss estimation in multilevel models by casting the multilevel model as a particular case of the general Bayes linear model and, hence, present estimates of β_j as posterior means of their corresponding posterior distribution.[5] Other approaches focus on the James-Stein "borrowing-of-strength" aspect of multilevel modeling when presenting estimates of the Level 1 coefficients.[6] Another alternative is to focus on the likelihood established by Equation 5, where maximum likelihood estimates for the three parameters σ^2, τ, and γ are obtained. Regardless, the main result is that the estimates of β_j may be

expressed as a linear combinations of the OLS estimate $\hat{\beta}_j = (\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{X}_j\mathbf{y}_j$ and—given an estimate of γ—the prior estimate $\mathbf{W}_j\hat{\gamma}$ of β_j, the weights being proportional to the estimation variance in the OLS estimate and the prior variance of the distribution of β_j. Thus, this may be viewed as a compromise between the within-group estimator, which ignores the data structure and the between-group estimator that models the within-group coefficients as varying around a conditional grand mean. More formally, assuming for now that the variance components and γ are known, the multilevel model estimate of β_j may be expressed as:

$$\hat{\beta}_j^* = \Theta_j\hat{\beta}_j + (\mathbf{I} - \Theta_j)\mathbf{W}_j\gamma, \tag{9}$$

where

$$\Theta_j = \tau\left(\tau + \sigma^2\left(\mathbf{X}_j'\,\mathbf{X}_j\right)^{-1}\right)^{-1} \tag{10}$$

is the ratio of the parameter variance for $\beta_j(\tau)$ relative to the variance $\sigma^2(\mathbf{X}_j'\mathbf{X}_j)^{-1}$ for the OLS estimator for β_j plus this parameter variance matrix. Thus, if the OLS estimate is unreliable, $\hat{\beta}_j^*$ will pull $\hat{\beta}_j$ toward $\mathbf{W}_j\hat{\gamma}$, the prior estimate.[7] Indeed, a little bit of algebra demonstrates that the shrinkage estimator in Equation 9 is the expected value of β_j given $\mathbf{y}_j$[8]:

$$\begin{aligned} E(\beta_j|\mathbf{y}_j) &= E(\beta_j) + \mathrm{Cov}(\beta_j, \mathbf{y}_j)[\mathrm{Var}(\mathbf{y}_j)]^{-1}[\mathbf{y}_j - E(\mathbf{y}_j)] \\ &= \mathbf{W}_j\gamma + \tau\mathbf{X}'_j\mathbf{V}_j^{-1}(\mathbf{y}_j - \mathbf{X}_j\mathbf{W}_j\gamma) \\ &= \mathbf{W}_j\gamma + \tau\mathbf{X}'_j\mathbf{V}_j^{-1}\mathbf{y}_j - \tau\mathbf{X}'_j\mathbf{V}_j^{-1}\mathbf{X}_j\mathbf{W}_j\gamma. \end{aligned} \tag{11}$$

Swamy (1971, p. 101) presents the following formula for the inverse of $\mathbf{V}_j$,

$$\mathbf{V}_j^{-1} = \sigma^{-2}\left[\mathbf{I} - \mathbf{X}_j(\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{X}_j'\right] + \mathbf{X}_j(\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{A}_j^{-1}(\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{X}_j', \tag{12}$$

where $\mathbf{A}_j = \tau + \sigma^2(\mathbf{X}_j'\mathbf{X}_j)^{-1}$. This implies that $\mathbf{X}_j'\mathbf{V}_j^{-1}\mathbf{X}_j = \mathbf{A}_j^{-1}$ and that $\mathbf{X}_j'\mathbf{V}_j^{-1}\mathbf{y}_j = \mathbf{A}_j^{-1}\hat{\beta}_j$ (de Leeuw & Kreft, 1986). Substituting these two results into the previous equation quickly leads to the desired result:

$$\begin{aligned} E(\beta_j|\,\mathbf{y}_j) &= \mathbf{W}_j\gamma + \tau\mathbf{A}_j^{-1}\hat{\beta}_j - \tau\mathbf{A}_j^{-1}\mathbf{W}_j\gamma \\ &= \tau\mathbf{A}_j^{-1}\hat{\beta}_j + \left(\mathbf{I} - \tau\mathbf{A}_j^{-1}\right)\mathbf{W}_j\gamma \\ &= \Theta_j\hat{\beta}_j + (\mathbf{I} - \Theta_j)\mathbf{W}_j\gamma. \end{aligned} \tag{13}$$

The conditional expectation representation of the shrinkage estimator is well-known as the minimum mean square linear estimator (MMSLE) of β_j (Chipman, 1964; Rao, 1965).[9]

One may also write the multilevel estimate as $\hat{\beta}_j^* = \mathbf{W}_j\gamma + \hat{\mathbf{u}}_j$, where we recall that u_j may be interpreted in the mixed model sense as the random effect of the jth group. From the literature on the estimation of random effects in mixed linear models, we have the commonly employed estimator of random effects:

$$\hat{\mathbf{u}}_j = \mathbf{C}_j^{-1}\mathbf{X}_j'(\mathbf{y}_j - \mathbf{X}_j\mathbf{W}_j\gamma), \tag{14}$$

where

$$\mathbf{C}_j = \mathbf{X}_j'\mathbf{X}_j + \sigma^2\tau. \tag{15}$$

The fixed effects γ are usually unknown and must be estimated. The estimation of the fixed effects is most easily discussed by ignoring the Level 1 β_j's altogether. In doing so, one focuses instead on the combined Equation 3, where the problem then becomes one of estimating the fixed effects γ in a mixed linear model, the result of which is the well known formula:

$$\gamma = \left(\sum_{j=1}^{J} \mathbf{W_j'}\,\mathbf{X_j'}\,\mathbf{V}_j^{-1}\mathbf{X}_j\mathbf{W}_j\right)^{-1} \sum_{j=1}^{J} \mathbf{W_j'}\,\mathbf{X_j'}\,\mathbf{V}_j^{-1}\mathbf{y}_j, \tag{16}$$

where

$$V_j = \mathrm{Var}(y_j) = \mathbf{X}_j\tau\mathbf{X}_j' + \sigma^2\mathbf{I}.$$

One may interpret the above estimator of γ as a generalized linear model (GLM) estimator. In the case of unknown γ, the shrinkage estimator of Equation 9 employing this estimator of γ yields the minimum mean square linear unbiased estimator (MMSLUE) of β_j (Harville, 1976).[10] Moreover, viewing the multilevel model as a Bayesian hierarchical model with flat priors on the variance components and γ, the Bayes rule for squared error loss is the posterior mean of the parameter (Berger, 1980, p. 161). For γ, the Bayes rule is the estimator in Equation 16.

de Leeuw and Kreft (1995) discuss alternative estimates of the fixed effects by means a two-step procedure, where one first obtains the OLS estimates of the β_j and then regresses these values on the $\mathbf{W}_j$ values. Regardless, this approach of focusing on the estimation of γ instead of β_j is preferred by some because we are actually estimating a parameter and do not want to risk blurring the distinction that β_j is a random variable. Furthermore, casting the multilevel model in the mixed model framework links multilevel model prediction to the more natural prediction problems that occur in such areas as repeated measures studies (see Rao, 1987).

The previous discussion assumes that the variance components are known. Although there is considerable agreement with respect to the estimation of fixed

effects, there is significantly less agreement with respect to the variance components. The maximum likelihood estimates of the variance components must be computed iteratively by means of procedures such as Fisher Scoring (Longford, 1988), iteratively reweighted generalized least squares (Goldstein, 1986), or the EM algorithm (Dempster, Laird, & Rubin, 1977). These and other procedures manifest themselves in several software packages: HLM (Raudenbush, Byrk, Cheong, & Congden, 2000), MIXOR (Hedeker & Gibbons, 1996), MLWIN (Rasbash et al., 2000), SAS Proc Mixed (Littell, Milliken, Stroup, & Wolfinger, 1996), and VARCL (Longford, 1988). In addition, the software package BUGS (Spriegelhalter, Thomas, & Best, 1996) incorporates fully Bayesian methods that have been introduced (Seltzer, 1993). Note, although Lindley and Smith (1972) provided a general framework for hierarchical data with complex error structures, the inability to estimate the covariance components for unbalanced data precluded using such models in practice. The introduction of the EM algorithm provided a numeric solution to this problem and paved the way to various other approaches mentioned above.

Although estimation in multilevel models is an important topic, it is not the focus of this article. The focus here lies in the prediction of a future observable y_{*j} and is elaborated in the next section.

2. Prediction in Multilevel Models

Prediction in multilevel models is considered in terms of forecasting unobserved (yet observable) units, either at Level 1 or Level 2. A concise definition is important because the potential for confusion arises from the close link of the multilevel model with the mixed linear model where one finds the term "prediction" reserved for estimating/predicting random effects.[11] Consider the school example. After carrying out a multilevel model analysis on some data, suppose we want to know the outcome ($\mathbf{y}$) for a student not in the data set. Formally, let y_{*j} be the unknown outcome measure, say mathematics score, for an unsampled student in the jth school, where school j is not necessarily in our sample or even known. Furthermore, let us assume that the multilevel model structure given above is true, although we know that the model is never true. The basic problem is to predict y_{*j}. We present three main approaches to the prediction of y_{*j} and examine their performance through a simulation study that extensively covers both the sample size and parameter space.

The three methods examined are multilevel prediction, prior prediction, and OLS prediction. These three predictive methods correspond to the three possible ways of estimating β_j for multilevel data discussed previously. The relative properties of these estimators is not of central interest because the focus is on the prediction of a future observable—estimation is a means to an end.[12] However, whether or not the results herein agree with multilevel studies on estimation is of interest. Guidelines exist for appropriately choosing the Level 1 and Level 2 sample sizes exist with respect to the estimation of fixed effects and variance components. (Bassiri, 1988; Busing, 1993; Kim 1990; Mok, 1995).

2.1. *OLS Prediction Method*

In this approach we emphasize that there is no Level 2 model, that is, the Level 1 β_j coefficients are not modeled as random variables regressed on Level 2 variables. Instead, there are simply J separate regression equations:

$$\mathbf{Y}_j = \mathbf{X}_j\beta_j + \mathbf{r}_j, \tag{17}$$

and, as before, the goal is to predict a future observation in the jth group, y_{*j}:

$$y_{*j} = \mathbf{X}_{*j}\beta_j + r_{*j}. \tag{18}$$

If y_{*j} were observed $\mathbf{X}_{*j}$ would merely represent a row of the $\mathbf{X}_j$ design matrix and that $r_{*j} \sim N(0, \sigma^2)$. For the prediction of y_{*j} one simply takes the OLS estimate estimate $\hat{\beta}_j$ obtained solely from the jth group and employs the following prediction rule:

$$\hat{y}_{*j} = \mathbf{X}_{*j}\hat{\beta}_j, \tag{19}$$

where

$$\hat{\beta}_j = (\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{X}_j\mathbf{y}_j.$$

Thus, despite the nested nature of the data and the fact that the assumption of a diagonal dispersion matrix is violated (recall that $\mathbf{V}_j = \mathbf{X}_j\tau\mathbf{X}_j + \sigma^2\mathbf{I}$), the conventional OLS procedure is used. There exists the risk of unstable prediction in the cases where the number of units within the groups is small and overfitting is a common problem for OLS. Nevertheless, there is the positive benefit of using a well known and more easily communicable statistical procedure.

2.2. *Prior Prediction Method*

In this case, the structure of the data is not ignored; instead, the setup of the Multilevel model is adopted. However, we stop short of an actual multilevel analysis, treating the Level 2 model equation as a prior for β_j and employing the estimate of that prior as the estimate for β_j. The technique for estimating γ will be that presented in Equation 16. Recalling that the multilevel estimate can be viewed as a weighted combination of the OLS estimate and the prior estimate, this approach corresponds to putting all of the weight on the prior. Hence, the prediction rule now becomes

$$\hat{y}_{*j} = \mathbf{X}_{*j}\mathbf{W}_j\hat{\gamma}, \tag{20}$$

where

$$\hat{\gamma} = \left(\sum_{j=1}^{J} \mathbf{W}_j'\mathbf{X}_j'\hat{\mathbf{V}}_j^{-1}\mathbf{X}_j\mathbf{W}_j\right)^{-1} \sum_{j=1}^{J} \mathbf{W}_j'\mathbf{X}_j'\hat{\mathbf{V}}_j^{-1}\mathbf{y}_j, \tag{21}$$

and

$$\hat{\mathbf{V}}_j = \hat{\mathrm{V}}\mathrm{ar}(\mathbf{y}_j) = \mathbf{X}_j\hat{\tau}\mathbf{X}_j' + \hat{\sigma}^2\mathbf{I},$$

where $\hat{\tau}$ and $\hat{\sigma}^2$ must be estimated iteratively via maximum likelihood. Consider the case when we do not have any Level 2 $\mathbf{W}_j$ information. In such a case, one may view all of the β_j as randomly varying around some mean level γ. Then in the prediction above the estimate of this mean level would be substituted for the estimate of each β_j. By introducing the Level 2 $\mathbf{W}_j$ information the concept of conditional exchangeability is being modeled, that is, given two schools with the same $\mathbf{W}_j$ information one expects their β_j to vary around the same mean level. To be sure, predictions based on a conditional grand mean will produce a much different prediction than that produced from the OLS method. However, it does use all the data and thus will not be vulnerable to small sample instability problems.

The prior prediction method may be viewed as a diagnostic check of the multilevel under consideration. Recall, the multilevel model is often used in an attempt to "borrow strength" in the James–Stein sense. Groups are modeled as conditionally exchangeable, and estimates are formed as weighted combinations of an ensemble estimate and a solo estimate, the ensemble being the prior and the solo being the OLS. If the Multilevel model under consideration is poor or incorrect, the "borrowing" of strength will not be a good idea, that is, the estimate should not be pulled toward the ensemble estimate and neither should any prediction. Hilden-Minton (1995) discusses this with respect to diagnostics and further developed Geisser's (1979) model criticism for the multilevel model.

2.3. Multilevel Prediction Method

In this case, the prediction rule is formed using the multilevel model estimate of β_j. Recall that this estimate may be written as follows:

$$\hat{\beta}_j^* = \Theta_j\hat{\beta}_j + (\mathbf{I} - \Theta_j)\mathbf{W}_j\hat{\gamma}, \tag{22}$$

where

$$\Theta_j = \tau\left(\tau + \sigma^2\left(\mathbf{X}_j'\mathbf{X}_j\right)^{-1}\right)^{-1} \tag{23}$$

is the ratio of the parameter variance for $\beta_j(\tau)$ relative to the variance for the OLS estimator for β_j plus this parameter variance matrix. Thus, if the OLS estimate is unreliable, $\hat{\beta}_j^*$ will pull $\hat{\beta}_j$ toward $\mathrm{W}_j\hat{\gamma}$, the prior estimate. With regard to the prediction rule, the multilevel estimate β_j is used to form the multilevel predictor:

$$y_{*j} = \mathbf{X}_{*j}\beta_j^*. \tag{24}$$

Given that the multilevel model estimate of the Level 1 coefficient is a shrinkage estimator, much of the multilevel literature revolves around the advantage of shrinkage estimators, how they borrow strength, and solve the instability of estimation problem along along with many other issues encountered when dealing with nested data. With respect to prediction, Gray, Goldstein, and Thomas (2001) consider the problem of predicting future "value-added" performance across groups from past trends. The main result is that such prediction is unreliable.[1] Rubin (1980) examined the performance of what he termed empirical Bayes predictors in his "Law School Validity Studies" article. His approach is similar to our approach of focusing on the predicting the future observable y_{*j}. However, his empirical Bayes predictor can be viewed as the basic multilevel model without any Level 2 variables. For his particular data set, he showed small gains using empirical Bayes predictors. However, he did not employ any Level 2 variables to extend his model to a full multilevel model. His searches for useful Level 2 variables to improve prediction failed to produce any viable candidates.[13] In addition to the advantage of shrinkage estimators, there is also a small literature on the dangers of shrinkage estimators, giving rise to limited translation rules (Efron & Morris, 1971, 1972). These represent safeguards for shrinking the estimators too far toward the ensemble estimate. The same concern exists for prediction, where predictions may be translated or shrunk too far, resulting in various practical worries.[14]

2.4. Analytical Results

In the Bayesian framework the model parameters are viewed as random variables. Thus, the prediction or "estimation" of the future observable y_{*j} is similar to the Bayesian estimation of a model parameter. The Bayes rule (under squared error loss) for predicting y_{*j} is the mean of the posterior distribution of y_{*j}, often known as the posterior predictive distribution. Employing the well known result that mean squared error (MSE) of an estimator is equal to the variance of the estimator plus the squared bias, the MSE of the estimator is the variance of the posterior predictive distribution. Thus, the estimator that will minimize the MSE is $E[y_{*j} \mid \mathbf{D}]$, where $\mathbf{D}$ is all the observable data. We now show that the mean of the posterior predictive distribution for an observation in group j is the estimator in Equation 24.

Proposition 1.1. *Assume the multilevel model defined by Equations 1 and 2 and that the variance components are known. Then the multilevel estimator of the Level 1 random coefficients is defined by Equation 9. Now suppose that we want to predict y_{*j}, that is, a future observation in group j. The mean of the posterior distribution of y_{*j} is the Multilevel prediction rule, that is, the Bayes rule under squared error loss is $\mathbf{X}_{*j}\hat{\beta}^{*}_{j}$.*

Proof. Because y_{*j} and Y are jointly Gaussian, we may write their joint distribution and employ well known results for conditional expectation and matrix identities to obtain the desired result. Formally, we have:

$$\begin{pmatrix} y_{*j} \\ \mathbf{Y} \end{pmatrix} \sim N\left[\begin{pmatrix} \mathbf{X}_{*j}\mathbf{W}_j\gamma \\ \mathbf{X}\mathbf{W}\gamma \end{pmatrix}, \begin{pmatrix} V_* & \mathbf{A} \\ \mathbf{A}' & \mathbf{X}\mathbf{T}\mathbf{X}' + \sigma^2\mathbf{I} \end{pmatrix}\right], \tag{25}$$

where

$$V_* = \mathrm{Var}(y_{*j}) = \mathbf{X}_{*j}\tau\mathbf{X}_{*j}' + \sigma^2$$
$$\mathbf{A} = \mathrm{Cov}(y_{*j}, \mathbf{Y}) = [0 \cdots 0\ \mathbf{V}_{*j}\ 0 \cdots 0].$$

For the row vector $\mathbf{A}$, each term 0 is of dimension $1 \times n_{j'}$ and represents $\mathrm{Cov}(y_{*j}, \mathbf{y}_{j'})$ for $j \neq j'$; and $\mathbf{V}_{*j} = \mathrm{Cov}(y_{*j}, \mathbf{y}_j) = \mathbf{X}_{*j}\tau\mathbf{X}_j'$. Because the distribution is a multivariate normal, we may directly write the conditional expectation of y_{*j} given the observable data $\mathbf{D}$, consisting of $\mathbf{Y}$ and $\mathbf{X}$:

$$E(y_{*j}|\mathbf{D}) = \mathbf{X}_{*j}\mathbf{W}_j\gamma + \mathbf{A}(\mathbf{X}\mathbf{T}\mathbf{X}' + \sigma^2\mathbf{I})^{-1}(\mathbf{Y} - \mathbf{X}\mathbf{W}\gamma).$$

Note that $\mathbf{X}\mathbf{T}\mathbf{X}' + \sigma^2\mathbf{I}$ is block diagonal, with each block equal to $\mathbf{V}_j = \mathbf{X}_j\tau\mathbf{X}_j' + \sigma^2\mathbf{I}$. Employing the structure of the matrix $\mathbf{A}$ and standard methods for inverting a symmetric partitioned matrix (see e.g., Morrison, 1967, p. 88), we obtain the following:

$$\begin{aligned} E(y_{*j}|\mathbf{Y}, \mathbf{X}) &= \mathbf{X}_{*j}\mathbf{W}_j\gamma + \mathbf{V}_{*j}\mathbf{V}_j^{-1}(\mathbf{Y}_j - \mathbf{X}_j\mathbf{W}_j\gamma) \\ &= \mathbf{X}_{*j}[\mathbf{W}_j\gamma + \tau\mathbf{X}_j'\mathbf{V}_j^{-1}(\mathbf{Y}_j - \mathbf{X}_j\mathbf{W}_j\gamma)] \\ &= \mathbf{X}_{*j}\hat{\beta}_j^*. \end{aligned}$$

The second equality follows from Equation 11 and the last equality follows from Equations 13 and 24, hence we have the desired result. □

In practice, the shrinkage estimator is often biased. Because the variance components are usually unknown, they are often estimated by means of the various maximum likelihood methods mentioned earlier. Although the MLEs have well behaved asymptotic properties, in small sample situations MLEs for the Level 2 variance components are frequently negatively biased (Afshartous, 1995; Busing, 1993; Raudenbush, 2003). Although the bias is reduced as the number of groups J increases, the likelihood for the Level 2 variances can exhibit significant positive skewness for small group sizes n_j, even when J is fairly large, thus the MLEs will still be negatively biased (Raudenbush, 2003). Because the shrinkage factor Θ_j in Equation 9 is conditional upon the estimates of the Level 2 variance components, any bias in the variance components will translate into a bias in $\hat{\beta}_j^*$. Regarding the amount of bias, consider the $p = 1$ case with no Level 1 covariates discussed by Morris (1995), thus the Level 2 variance τ is now a scalar. Assume that τ is underestimated by a factor υ, that is, $\hat{\tau}$ has expectation $\upsilon\tau$. Without loss of generality, assume σ^2 is known.[15] A decrease in the estimate for τ directly results in an increase in the weight accorded to the prior estimate, that is, more shrinkage. The expected amount of this increase in shrinkage is a function of the Level 1 variance, Level 2 variance, and υ. Formally, we have:

$$\text{increased shrinkage} = f(\sigma^2, \tau, \upsilon) = \frac{(1 - \upsilon)\tau}{\sigma^2 + \upsilon\tau}. \tag{26}$$

The expression above may be viewed as the expected increase in shrinkage when using a biased estimator of τ versus an unbiased estimator of τ. For example, suppose $\upsilon = .9$, that is, we underestimate the Level 2 variance by 10%. If $\tau = 1$, this underestimation of the Level 2 variance results in 47% additional shrinkage, on average. Hence, the resulting estimator for the random Level 1 coefficient will be biased. The amount of this bias is dependent on the particular Level 1 and Level 2 model and the values of the corresponding design matrices for the full model. Browne and Draper (2002) have employed MCMC methods such as Gibbs sampling and Metropolis–Hasting algorithms to obtain reduced bias in variance component estimation; to offset the drawback of slower computational speed of the MCMC methods, they propose a hybrid approach based on likelihood and MCMC methods.[16]

Our original prediction problem is slightly altered if we seek to predict the outcome for an individual from a group not in the sample; that is, $y_{*j'}$ where $j' \neq j$, $\forall j = 1, \ldots, J$. In Equation 25, we must replace y_{*j} by $y_{*j'}$ accordingly. The off-diagonal blocks of the variance of the joint distribution are now null because $\mathrm{Cov}(y_{*j'}, \mathbf{Y}) = 0$. Once again using standard results from multivariate normal theory, the expected value and variance for the posterior predictive distribution of $y_{*j'}$ given all the observable data $\mathbf{D}$ are $E(y_{*j'} \mid \mathbf{D}) = \mathbf{X}_{*j}'\mathbf{W}_j'\gamma$ and $\mathrm{Var}(y_{*j'} \mid \mathbf{D}) = \mathbf{X}_{*j'}\tau\mathbf{X}_{*j'}'$, respectively. Because MSE is equal to the variance of the posterior predictive distribution, the MSE will be relatively unaffected by the size of n_j, and the number of groups J will be more important. This follows from the well known result that the estimator for τ improves as the number of groups J increases.

The multilevel prediction rule of Equation 24 and the OLS prediction rule of Equation 19 use the multilevel (shrinkage) estimator and OLS estimator for the Level 1 regression coefficients, respectively. As the amount of data in each group increases, one would expect the multilevel predictions to become more similar to the OLS predictions because there will be less shrinkage away from the OLS estimator. We now formally demonstrate that the multilevel estimator approaches the OLS estimator as the group sizes n_j become large. Because the aforementioned prediction rules are linear functions of the estimators, it will follow that the multilevel prediction rule will approach the OLS prediction rule as n_j grows, thus the MSEs will become more similar as well.

Proposition 1.2. *Assume the multilevel model defined by Equations 1 and 2. Let the multilevel estimator for the Level 1 regression coefficients be defined by Equations 9 and 11; let the corresponding OLS estimator be defined by usual estimator* $\hat{\beta}_j = (X_j'\mathbf{X}_j)^{-1}\mathbf{X}_j\mathbf{y}_j$. *We have the following result:*

$$n_j \to \infty \Rightarrow \beta_j^* \to \hat{\beta}_j. \tag{27}$$

Proof. Recall that $\hat{\beta}_j^*$ may be written as a linear combination of β_j and a prior estimate. Given the form of this linear combination, we need to show that $\Theta_j \to \mathbf{I}$ as $n_j \to \infty$. Recall that the expression for Θ_j may be written as follows:

$$\Theta_j = \tau\left(\tau + \sigma^2\left(\mathbf{X}_j'\mathbf{X}_j\right)^{-1}\right)^{-1}. \tag{28}$$

Thus, we need to demonstrate that as $n_j \rightarrow \infty$, it follows that $\mathbf{X}_j'\mathbf{X}_j \rightarrow \infty$, where ∞ represents a $p \times p$ matrix with infinitely large elements. This will imply that $\Theta_j \rightarrow \tau(\tau)^{-1} = \mathbf{I}$, and we will have the desired result. Without loss of generality assume that $p = 2$. Thus, we have

$$\mathbf{X}_j'\mathbf{X}_j = \begin{pmatrix} n_j & \sum_{i=1}^{n_j} X_{ij} \\ \sum_{i=1}^{n_j} X_{ij} & \sum_{i=1}^{n_j} X^2{}_{ij} \end{pmatrix}.$$

Because n_j is an integer sequence, the upper-left term of the matrix above is clearly unbounded. The off-diagonal terms represent the sum of the predictor variables in group j, and this will be unbounded unless for every $\epsilon > 0$, there exists an N such that $k \geq N$ implies $|X_{kj} + X_{k+1,j} + \cdots + X_{k+p,j}| < \epsilon$, $\forall p = 1, 2, \ldots$ (see, e.g., Marsden & Hoffman, 1974, p. 135). A similar criterion may be applied to the lower-right term in the matrix. Thus, as long as the sum of predictors and sum of squared predictors are not convergent series as $n_j \rightarrow \infty$, we have the desired result. In most cases with real data, this will not be a stringent constraint. □

Raudenbush (2003) suggests that when considering the random effects u_j as missing data and $\mathbf{y}_j$ as the observed data, we can define $\mathbf{I} - \Theta_j$ as the fraction of missing information in cluster j. Thus, if this fraction is small, we shrink the multilevel estimator more toward the OLS estimator, as mentioned above.

Next, we demonstrate that the MSE of the multilevel prediction rule approaches the MSE of the prior prediction rule as the intraclass correlation goes to zero. The intraclass correlation, defined as $\rho = \dfrac{\tau_{00}}{\tau_{00} + \sigma^2}$, measures the degree to which observations within the same group are related. When ρ is large, the groups are relatively homogeneous with large variability across groups, while when ρ is small, there is large variability within groups as opposed to between groups. Because the multilevel and prior prediction rules are linear functions of their respective estimators for the Level 1 regression coefficients, we will have the desired result by proving the following proposition.

Proposition 1.3. *Assume the multilevel model defined by Equations 1 and 2. Let the multilevel estimator for the Level 1 regression coefficients be defined by Equations 9 and 11; let the corresponding prior estimator be defined by* $\hat{\beta}_j^{\text{Prior}} = \mathbf{W}_j\hat{\gamma}$, *where* $\hat{\gamma}$ *is defined as in Equation 21. Let the intraclass correlation coefficient be defined as* $\rho = \dfrac{\tau_{00}}{\tau_{00} + \sigma^2}$. *We then have the following result:*

$$\rho \rightarrow 0 \Rightarrow \beta_j^* \rightarrow \hat{\beta}_j^{\text{Prior}}. \tag{29}$$

Proof. Without loss of generality assume that γ is known and that $p = 2$. As in the previous proof, we focus on the shrinkage format of the multilevel estimator given in Equation 9. Notice that the result will immediately follow if we can show that

$\rho \to 0$ implies that $\Theta_j \to 0$, where 0 represents a 2×2 zero-matrix. For the $p = 2$ case, assuming arbitrary constants a, b, c, d for the elements of $(\mathbf{X}_j'\mathbf{X}_j)^{-1}$, we have:

$$
\begin{aligned}
\Theta &= \tau\left(\tau + \sigma^2\left(\mathbf{X}_j'\mathbf{X}_j\right)^{-1}\right)^{-1} \\
&= \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{pmatrix}\begin{pmatrix} \tau_{00} + \sigma^2 a & \tau_{01} + \sigma^2 c \\ \tau_{10} + \sigma^2 b & \tau_{11} + \sigma^2 d \end{pmatrix}^{-1} \\
&= \xi\begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{pmatrix}\begin{pmatrix} \tau_{11} + \sigma^2 d & -\tau_{01} - \sigma^2 c \\ -\tau_{10} - \sigma^2 b & \tau_{00} + \sigma^2 a \end{pmatrix} \\
&= \xi\begin{pmatrix} \tau_{00}\tau_{11} + \tau_{00}\sigma^2 - \tau_{01^2} - \tau_{01}\sigma^2 & -\tau_{00}\tau_{01} - \tau_{00}\sigma^2 - \tau_{01}\tau_{00} + \tau_{01}\sigma^2 \\ \tau_{10}\tau_{11} + \tau_{10}\sigma^2 - \tau_{10}\tau_{11} - \tau_{11}\sigma^2 & -\tau_{10}\tau_{01} - \tau_{10}\sigma^2 + \tau_{11}\tau_{00} + \tau_{11}\sigma^2 \end{pmatrix},
\end{aligned}
$$

where $\xi = [(\tau_{00} + \sigma^2 a)(\tau_{11} + \sigma^2 d) - (\tau_{10} + \sigma^2 b)(\tau_{01} + \sigma^2 c)]^{-1}$, and the third equality follows from the general rule for inverting a 2×2 matrix (see, e.g., Strang, 1988). By assumption $\rho \to 0$, which implies that $\sigma \to \infty$. In the last line above, each term in the matrix is of order σ^2, while ξ has terms in its denominator of order σ^4. Thus, multiplying the matrix term by term by ξ, it follows that each matrix term converges to 0 as $\sigma^2 \to \infty$ by applying l'Hospital's rule. Hence we have the desired result. □

When the groups are more similar, it makes sense to "borrow strength" from other groups by means of a prior estimator, whereas if the variation between groups is high, employing the prior estimator is potentially dangerous for a given group. Indeed, for the special case of a simple hierarchical model with $p = 1$ and no Level 1 covariates, the shrinkage factor Θ_j is a scalar and is *equal* to ρ. Thus, given the form of Equation 9, it is clear that as ρ approaches 0, the estimator is shrunk more toward the prior estimator.[17]

Although the analytical results above are valuable demonstrations of the behavior of the estimators and prediction rules, it is often useful to complement these results with simulations that cover both a wide sample space (Level 1 and Level 2) as well as a wide parameter design space. This will allow the side-by-side comparison of different areas of the sample and parameter design space in a clear manner. For example, these simulations will provide guidelines on the areas in the sample and parameter design space in which the multilevel prediction rule is most beneficial. The relative performance of all three prediction rules is assessed via an extensive simulation study, which is described in the next section.

3. Simulation Study Design

Multilevel data are simulated under a variety of design conditions, closely following the simulation study of Busing (1993), where the distribution of Level 2 variance component estimates was examined. As in Busing, a simple 2-level multilevel model with one explanatory variable at each level and equal numbers of units per group is considered. A two-stage simulation scheme is employed. At the first stage, the Level 1 random coefficients are generated according to the following equations[18]:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j},$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11} W_j + u_{1j}.$$

The γs are the fixed effects and are set to a predetermined value; they are set all equal to one as in Busing. W_j is a standard normal random variable, while the error components, u_{0j} and u_{1j}, have a bivariate normal distribution with mean (0, 0) and a 2×2 covariance matrix τ. The two diagonal elements of τ, τ_{00} and τ_{11}, are equal in each design condition. The off-diagonal covariance term τ_{01} will then determine the correlation between the intercept and slope:

$$r_{u_{0j}, u_{1j}} = \frac{\tau_{01}}{(\tau_{00}\tau_{11})^{1/2}}. \tag{30}$$

Another parameter of interest in the simulation design is the intraclass correlation ρ. The intraclass correlation is defined as follows:

$$\rho = \frac{\tau_{00}}{\tau_{00} + \sigma^2}, \tag{31}$$

and thus measures the degree to which units within the same unit are related. Intraclass correlations of 0.2 and above are common in educational research; a range of intraclass values of 0.2, 0.4, 0.6, and 0.8 is examined to provide information for both high and low intraclass correlation conditions.

The second stage of the simulation concerns the first level of the multilevel model, where observations are generated according to the following equation:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + \epsilon_{ij}. \tag{32}$$

The Level 2 outcome variables, the βs, were determined at the first stage of the simulation. The Level 1 explanatory variable X_{ij} is simulated as a standard normal random variable, while the Level 1 error ϵ_{ij} is a normal random variable with mean 0 and variance σ^2 specified as 0.5. Because only the balanced data case is considered, where there are n units grouped within J groups, a total of Jn outcomes are simulated. To study prediction, an extra $(n+1)$st observation is simulated for each of the J groups; this observation is set aside and is not used for estimative purposes; this is the future observable y_{*j} for which the prediction rules are applied. Table 1 and Table 2 summarize the various parameter specifications in the simulation design.

Simulations are conducted under various sample size combinations for the number of groups (J) and the number of observations per group (n). Information concerning the effects of J and n with respect to the performance of prediction rules is of practical interest at the design or data-gathering phase. Given one's research interests, one would want to know the appropriate values for the number of groups and number of elements per group to sample, especially given the increased cost of

TABLE 1
ρ, τ_{00}, τ_{11}

Intraclass correlation ρ	0.200	0.400	0.600	0.800
Variance τ_{00}, τ_{11}	0.125	0.333	0.750	2.00

TABLE 2
τ_{01}

	Correlation Intercepts-Slopes		
Variance	0.25000	0.5000	0.75000
0.125	0.03125	0.0625	0.09375
0.333	0.08330	0.1667	0.25000
0.75	0.18750	0.3750	0.56250
2.0	0.50000	1.0000	1.50000

including an additional group in one's study. Thus, an extensive sample size space is explored in this simulation study. The layout of the design is given in Table 3.

The simulation design represents a three-factor full factorial design, where factor one is Level 1 sample size (five values), factor two is Level 2 sample size (five values), and factor three is parameter values (12 values), yielding a total of 300 design conditions. As mentioned above, one additional observation per group is simulated, which is used to assess the prediction rules. Thus, when $J = 10$, there will be 10 predictions for a given data set. In addition, for each design condition, 100 replications are performed, that is, 100 multilevel data sets are simulated for each design condition, and prediction is assessed within each of these replications. Thus, because there are 300 design conditions, a total of 30,000 multilevel data sets will be generated in this initial part of the study.

This next phase of this simulation study represents a comparison of the three predictors presented earlier: multilevel, prior, and OLS. Recall that the goal is to predict a future observable y_{*j} in each of the J groups and replicate this process 100 times to account for variability. The adequacy of prediction is measured via predictive

TABLE 3
Sample Sizes

	n_j				
J	5	10	25	50	100
10	50	100	250	500	1,000
25	125	250	625	1,250	2,500
50	250	500	1,250	2,500	5,000
100	500	1,000	2,500	5,000	10,000
300	1,000	3,000	7,500	15,000	30,000

TABLE 4
Design Numbers

Design Number	τ_{00}, τ_{11}	τ_{01}	$r_{u_{0j}, u_{1j}}$	ρ
1	0.125	0.03125	0.25000	0.200
2	0.333	0.08330	0.25000	0.400
3	0.75	0.1875	0.25000	0.600
4	2.0	0.50000	0.25000	0.800
5	0.125	0.0625	0.5000	0.200
6	0.333	0.1667	0.5000	0.400
7	0.75	0.3750	0.5000	0.600
8	2.0	1.0000	0.5000	0.800
9	0.125	0.09375	0.75000	0.200
10	0.333	0.25000	0.75000	0.400
11	0.75	0.56250	0.75000	0.600
12	2.0	1.50000	0.75000	0.800

mean square error (PMSE), where the popular technique of taking the average of the sum the squared errors (SSE) of the observed and predicted values is employed.[19] Thus, for each of the 300 design conditions there are 100 replications of the predictive mean square error for each prediction rule. Note that this PMSE is constructed from a different number of items in the different sample size combinations. For instance, when $J = 10$ each replication consists of predicting 10 future observables, thus the PMSE is the average of 10 squared difference, while for $J = 300$ each replication consists of predicting 300 future observables, thus the PMSE is the average of 300 squared differences. Because 100 replications are taken, the average of PMSE over the replications should be fairly reliable and enable the comparison across design conditions for variability in PMSE.

Table 4 summarizes the combinations of parameter values for the simulation study. Because of space considerations, results are mainly presented for design conditions with low ($\rho = 0.2$) and high ($\rho = 0.8$) intraclass correlations.[20]

3.1. Terrace-Two

The computer code for generating the data was written in XLISP-STAT,[21] and the multilevel modeling was done with several altered versions of Terrace-Two.[22] Although many of the more popular multilevel software packages are faster, the object oriented nature of XLISP-STAT facilitated the amendment and alteration of Terrace-Two in order to extend its capability. Defaults such as the maximum number of iterations were changed to allow the number of replications to proceed in the background. Regarding computing time, some of the higher level $J \times n$ sample size combinations were very computer intensive, requiring several hours of computing time on Sun Sparc 10 machines. The limiting factor in the simulations was the actual estimation of the multilevel model, which is a function of J, the number of groups, and not $N = Jn$, the total sample sizes. The data simulations and formation of prediction rules after estimation required very little computing time.

4. Results

The simulation results for all parametric designs clearly indicate that the multilevel method consistently produces the lowest PMSE across each of the $J \times n$ sample size combinations. Specifically, the Multilevel prediction rule produced the lowest average PMSE in 24 of the 25 possible $J \times n$ combinations, the only exception being the $J = 10$, $n = 50$ case where the OLS prediction rule produced a nearly identical PMSE to that of the multilevel prediction rule (0.2640 vs. 0.2651, respectively). As expected, the differential in PMSE between the multilevel and OLS prediction rules becomes less as the group size n increases, a result of the increased reliability of the OLS prediction in such cases. Note that an increase in the number of groups should have little if any effect on the OLS prediction rule because this method produces prediction independently in each group. As the group size n increases, however, the OLS prediction rule produces PMSEs very similar to that of the multilevel rule, albeit consistently higher.

The prior prediction rule consistently performs the worst of the three methods, and very much so in absolute terms, more than a full unit higher in PMSE in all $J \times n$ combinations. Although increasing the group size n has little effect on the prior prediction rule, there is a considerable rise in PMSE for the prior prediction rule as the number of groups J rises. Recall that the prior prediction method uses the rule $\hat{y}_{*j} = \mathbf{W}_j\hat{\gamma}$. Bassiri (1988) demonstrated that an increase in J is beneficial with respect to estimation of γ, while here it seems that the prior prediction rule—the performance of which solely depends on the estimation of γ—performs worse when J is increased. Although this may seem contradictory, it is a manifestation of the dangers of using a grand mean to predict at the individual level. For instance, the estimate of $\hat{\gamma}$ is formed by means of Equation 16, which is a sum over J groups. For small J values, this would be fairly representative of the space of groups, whereas for large J, this would be less so because the sum would involve many more terms, each sum with its own values for group specific information such as $\mathbf{V}_j$. Thus, as J increases, the chances of mispredicting within a particular group increases, leading to the exhibited behavior of the prior prediction rule.

The aforementioned results are illustrated by means of the three-dimensional plots in Figures 1–3, which cover the parametric designs with an intraclass correlation of 0.2. Figure 1 illustrates the PMSE of the multilevel prediction rule, while Figure 2 presents the difference between PMSE from the OLS to the multilevel prediction rule, and likewise, for the difference between prior and multilevel prediction rules in Figure 3. Figure 2 illustrates that the advantage of the multilevel prediction rule over the OLS prediction rule is clearly best for low values of n, for example, $n = 5$ and $n = 10$.[23] Figures 1 and 2 illustrate the improved PMSE as group size n increases for both the multilevel and OLS prediction rules, for all levels of J. In addition, the narrowing of the differential between the multilevel and OLS prediction rules as n increases is also clear for each level of J. Figure 3 illustrates the large difference in PMSE between the prior and multilevel prediction rules, in addition to the adverse effect of an increase in J with respect to the PMSE for the prior prediction rule.[24] With respect to the multilevel prediction rule, although the

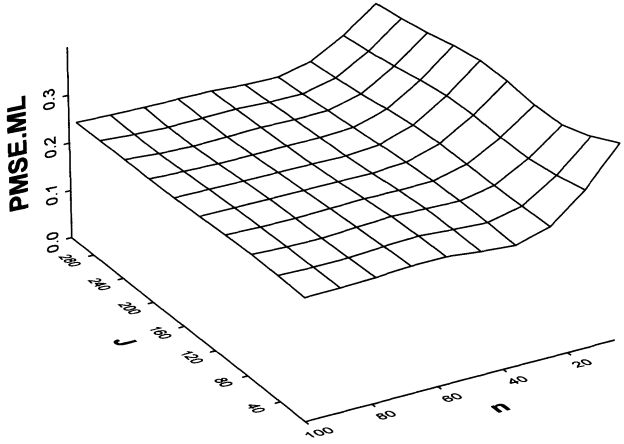

FIGURE 1. *PMSE for multilevel prediction rule,* $\rho = 0.2$.

reduction in PMSE as J increases is slight, the overall simulation results demonstrate that there is a reduction in the variability of PMSE as J increases. For the prior prediction rule, however, not only does the average level of PMSE increase as J increases, the variability in PMSE increases as well.

The corresponding three-dimensional plots of PMSE for a high intraclass correlation ($\rho = 0.8$) are shown in Figures 4–6.[25] Similar patterns are evident with respect to the effect on PMSE for changes in n and J, albeit there is a slight upward shift in PMSE for all three prediction methods. For the multilevel prediction rule,

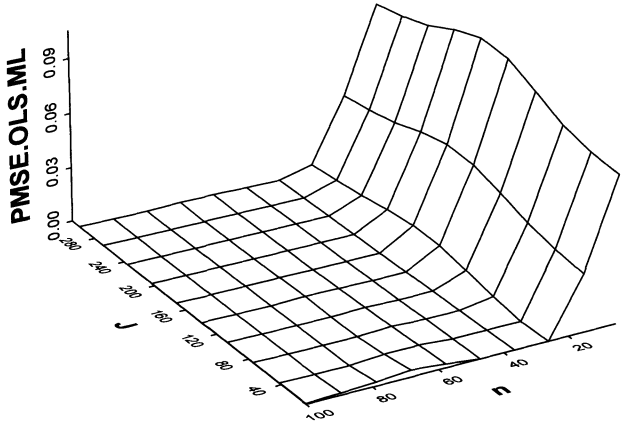

FIGURE 2. *Differential PMSE: OLS minus multilevel PMSE,* $\rho = 0.2$.

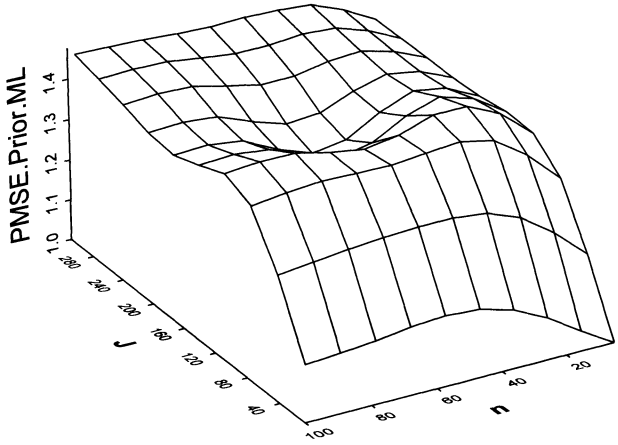

FIGURE 3. *Differential PMSE: Prior minus multilevel PMSE,* $\rho = 0.2$.

this shift is greatest for small values of n. For example, for $J = 100$ and $n = 5$, the PMSEs for the $\rho = 0.2$ and $\rho = 0.8$ designs are 0.36 and 0.41, respectively. The difference between the multilevel and OLS PMSE is reduced under this higher intraclass correlation. For example, for $J = 100$ and $n = 5$, the differential PMSE decreases from 0.086 to 0.03, mainly because of the poorer performance of the multilevel prediction rule under high intraclass correlation. The prior prediction rule performs extremely poorly under this higher intraclass correlation, thereby resulting in a larger

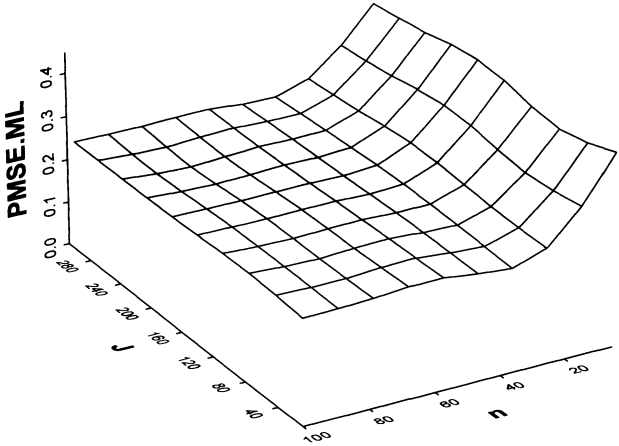

FIGURE 4. *PMSE for multilevel prediction rule,* $\rho = 0.8$.

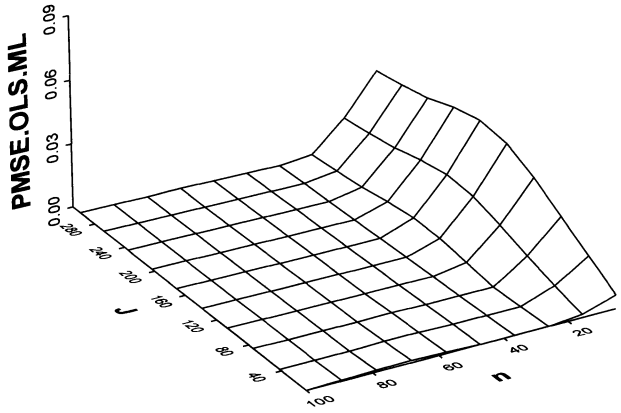

FIGURE 5. *Differential PMSE: OLS minus multilevel PMSE,* $\rho = 0.8$.

difference between PMSE for the multilevel and prior prediction rules, for example, almost tripling the differential PMSE from 1.35 to 3.64 for $J = 100$ and $n = 5$.

However, for intraclass correlation coefficients of 0.2 to 0.6, the prior method is not as far off the multilevel and OLS prediction rules, although its ranking is still a consistent third across all combinations of J and n. The results for the prior method cluster into two distinct groups, that is, those for intraclass correlation values of 0.6 and lower and those for an intraclass correlation value of 0.8. On the other hand, the multilevel and OLS prediction rules exhibit an even distribution across the differ-

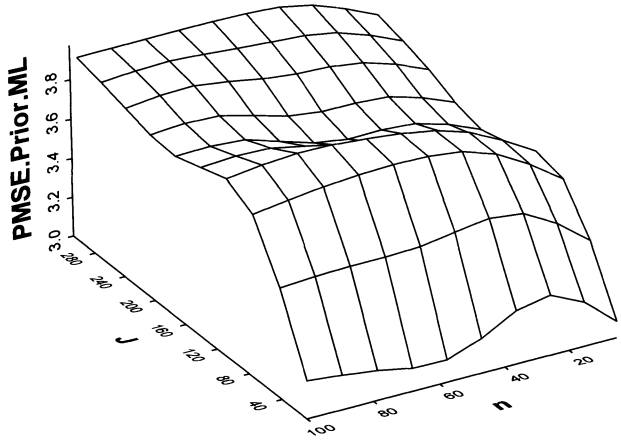

FIGURE 6. *Differential PMSE: Prior minus multilevel PMSE,* $\rho = 0.8$.

ent values of the intraclass correlation. Figure 7 illustrates these results by means of a scatterplot matrix of PMSE for one particular combination of $J \times n$: $J = 25$ and $n = 5$. Each point represents the average PMSE taken over 100 replications for each parametric design condition.

The overall results indicate that a predictive perspective often leads to decisions that differ from those arising from an estimative perspective. Specifically, the results indicate that an increase in group size n is often more beneficial with respect to prediction than an increase in the number of groups J. With respect to the estimation of multilevel model parameters, previous simulation studies (Bassiri, 1988; Busing, 1993; Mok, 1995) indicate that estimation is more improved by increasing the number of groups J instead of the group size n. For example, the sampling distribution of $\hat{\tau}$ is skewed to the right with the true value to the right of the mean of this skewed distribution. Thus, $\hat{\tau}$ will be negatively biased in estimating τ. Busing (1994) demonstrated that an increase in the number of groups J was beneficial in reducing this bias, whereas an increase in group size (n) had no affect. With respect to $\hat{\sigma}^2$, Busing obtained relative bias close to zero for all sample design conditions except the smallest sample sizes. This is because these Level 1 variance estimates are based on the total sample size N, thus this estimate will show little bias as N is large. Finally, focusing on the fixed effects instead of the variance components, Bassiri (1988) demonstrated the improved estimation of γ as the number of groups J increases.

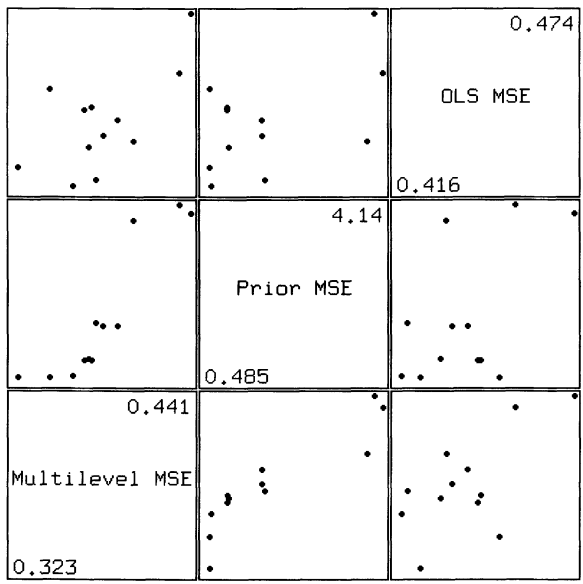

FIGURE 7. *Scatterplot matrix of predictive MSE; $J = 25$, $n = 5$.*

4.1. *Model Misspecification*

The results above do not account for model uncertainty because the correct model is estimated when forming predictions. In practice, the multilevel model may be incorrect in a variety of ways. For example, the Level 2 model may be misspecified for either the Level 1 intercept or slope(s) by failing to include the appropriate Level 2 covariates. On the other hand, the full multilevel model that correctly includes Level 2 variables may be fitted, only to have the assumptions of that model violated, for example: (a) the error term at Level 1 and/or Level 2 may be Cauchy or t distributed rather than normal, (b) the Level 1 observations (students) within each Level 2 unit (school) may be dependent, and (c) the Level 2 units may also be dependent. We limit ourselves, however, to studying Level 2 variable misspecification; the latter error and independence violations are the subject of future research.

To examine the effect of Level 2 model misspecification on prediction, we consider three specific misspecifications, all concerning the Level 2 equations that model the random Level 1 coefficients. First, the model may be misspecified by failing to include the Level 2 variable when modeling the random Level 1 slope:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + u_{1j}.$$

Note, this is not to say that such model specifications are not sometimes useful or "correct," but rather that in the current study, where we have employed simulation based on a model that does indeed have a Level 2 variable as part of the stochastic process that generates the Level 1 slope, it would be a misspecification were we not to include it in the modeling of the Level 1 slope. Second, the model may be misspecified by failing to include the Level 2 variable when modeling the random Level 1 intercept:

$$\beta_{0j} = \gamma_{00} + u_{0j},$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11} W_j + u_{1j}.$$

A reasonable question to ask is the following: Is the penalty for omitting the Level 2 variable symmetric with respect to the random slope and intercept? If it is not, one would obviously want to devote more time to the modeling of the Level 2 equation where the penalty is greater. Finally, the model may be misspecified by failing to include Level 2 variable altogether in the Level 2 coefficient equations:

$$\beta_{0j} = \gamma_{00} + u_{0j},$$
$$\beta_{1j} = \gamma_{10} + u_{1j}.$$

TABLE 5
Misspecification Type

Name	Type of Misspecification
Misspec A	W_j used only for the slope
Misspec B	W_j used only for the intercept
Misspec C	W_j not used in Level 2 equation

This formulation represents a random intercept model: the requirement of conditional exchangeability is no longer necessary, that is, the Level 1 random coefficients are assumed (incorrectly) to vary around a single grand mean.

These three misspecifications are investigated by means of a simulation study following the design of the previous section. Given the poor relative performance of the prior prediction rule, we only compare the multilevel and OLS prediction rules. Because the OLS prediction rule is independent of Level 2 model misspecification, we seek to determine if the multilevel prediction rule still outperforms OLS, even when the model used to form the predictions is misspecified. Each design condition will report the average PMSEs over 100 replications for each of the three misspecifications. Table 5 summarizes the three misspecifications.

For values of n greater than five, the performance of the multilevel prediction rule is similar for all three misspecifications, and its relative performance to that of the OLS prediction rule is essentially unchanged from earlier. Table 6 contains the results obtained by averaging over the 12 parameter design conditions for the $n = 5$ case. For each value of J, misspecification B, that is, failing to include the Level 2 variable for the slope, produces the highest PMSE for the multilevel prediction rule, even higher than misspecification C, where the Level 2 variable is omitted from the model altogether. Although the results are similar in magnitude, misspecification A (Level 2 variable used only for slope) outperforms misspecification C (Level 2 variable omitted) slightly for three of the four values of J in Table 6. In addition, the Multilevel prediction rule under all three misspecifications produces a lower PMSE than that produced by the OLS prediction rule. Thus, it is encouraging that the previous results do not get reversed because of minor misspecifications of the model. Moreover, the PMSEs for misspecification A and C are very close to that of the correctly

TABLE 6
Mean MSE for Multilevel Misspec A, B, C, and OLS

	$n = 5$			
J	Misspec A	Misspec B	Misspec C	OLS
10	0.4082	0.4200	0.4019	0.4359
25	0.3960	0.4073	0.3982	0.4533
50	0.3948	0.4040	0.3967	0.4527
100	0.3868	0.3977	0.3905	0.4502

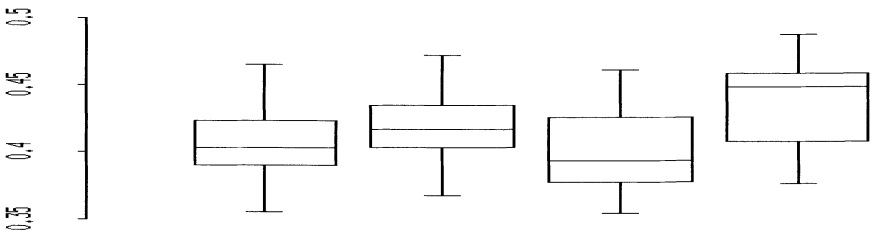

FIGURE 8. *J = 10; n = 5; MSE for multilevel misspec A, B, C, and OLS.*

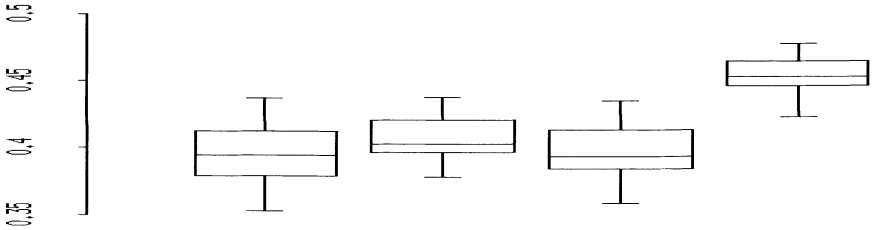

FIGURE 9. *J = 10; n = 25; MSE for multilevel misspec A, B, C, and OLS.*

specified model. At the same time, the consistently poor results of misspecification B relative to the other misspecifications indicates that one should pay close attention to the modeling of the random slope in the Level 2 equation because the omission of relevant Level 2 explanatory variables is more costly when modeling the Level 1 slope than when modeling the Level 1 intercept. Figures 8–11 illustrate this clearly.

5. Summary

In summary, we have advocated a predictive approach to multilevel modeling in which the focus lies on the prediction of future observables instead of the characteristics of estimators. Of course, because the prediction rules (OLS, prior, and multilevel) are in one-to-one correspondence to the estimation methods of β_j, these

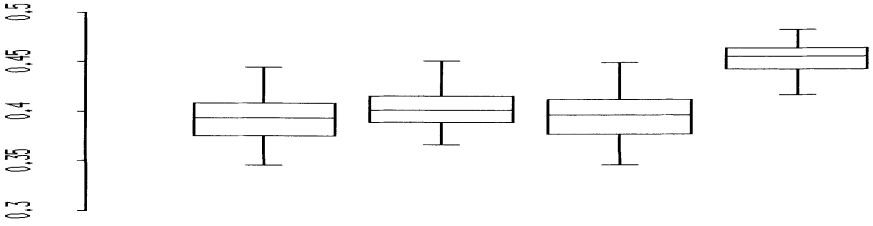

FIGURE 10. *J = 50; n = 5; MSE for multilevel misspec A, B, C, and OLS.*

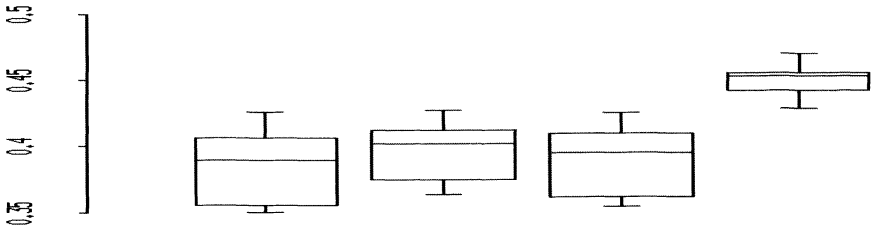

FIGURE 11. *$J = 100$; $n = 5$; MSE for multilevel misspec A, B, C, and OLS.*

two areas are related. We have presented both analytical and Monte Carlo results to assess the relative performance of the prediction rules. The analytical results are: (a) The multilevel prediction rule employing the shrinkage estimator for β_j is the Bayes rule for minimizing squared error loss and is the mean of the posterior distribution of y_{*j}; (b) bias in estimating the Level 2 variance translates into bias in the multilevel estimator, hence bias in the multilevel prediction rule; this bias may be represented via a function that models the change in shrinkage caused by bias in estimation; (c) when predicting a future observable for a group not in the sample, the number of groups becomes more important compared to the number of observations per group; (d) as group size becomes large, the multilevel estimator and prediction rule approach the OLS estimator and prediction rule, respectively; and (e) as the intraclass correlation ρ approaches 0, the multilevel and prior prediction rules become more similar with respect to MSE.

In the simulations, we observe that a predictive perspective often leads to decisions that differ from those arising from an estimative perspective. Specifically, the results indicate that an increase in group size n is often more beneficial with respect to prediction than an increase in the number of groups J. With respect to the estimation of multilevel model parameters, previous simulation studies (Bassiri, 1988; Busing, 1993; Mok, 1995) indicate that estimation is improved more by increasing the number of groups instead of the group size.

The simulations show that the multilevel prediction rule performed the best with respect to MSE across the various design specifications. This is to be expected based on the analytical demonstration that the multilevel estimator is the Bayes rule under squared error loss. The simulations also suggest that the predictive ability of the prior prediction rule is not only poor, but also weakens as the number of groups J increases. Although the former should be clear from the analytical results, the latter became more vivid during the simulations. Finally, although the results indicate that the multilevel prediction rule is fairly robust with respect to misspecification of the Level 2 model, there seems to be a differential cost with respect to misspecification of the random slope versus the random intercept. The main results for the simulations are further summarized below.

1. The multilevel prediction rule is clearly the best across the $J \times n$ combinations. This readily follows from Proposition 1.1.

2. PMSE is reduced as group size n increases for both the multilevel and OLS prediction rules, for all levels of J. Note that this is for prediction of a future observable that belongs to one of the groups in the sample. Our analytical results demonstrate that this would not be the case if the future observable did not belong to a group in the sample; rather, the number of groups would be more important.

3. The differential in PMSE between the multilevel and OLS prediction rules becomes less as the group size n increases. This result is expected from Proposition 1.2.

4. The prior prediction rule consistently performs the worst of the three prediction rules in absolute terms—more than a full unit higher in PMSE in all $J \times n$ combinations. However, the performance of the multilevel and prior prediction rules are similar as the intraclass correlation ρ approaches 0, as expected from Proposition 1.3.

5. There is an adverse effect of an increase in J with respect to the PMSE of the prior prediction rule. Not only does the average level of PMSE increase as J increases, the variability in PMSE also increases as well. These results are also dependent upon the influence of the value of ρ mentioned above.

6. With respect to the multilevel prediction rule, there is only a slight reduction in the overall level of PMSE as J increases. However, there is a reduction in the variability of PMSE for the multilevel prediction rule as J increases. Once again we note the analytical result that the number of groups J would have greater influence on PMSE if we were predicting a future observable that does not belong to one of the groups in the sample.

7. While the multilevel and OLS prediction rules exhibit a relatively even and narrow distribution across the 12 parameter conditions, such is not the case for the prior prediction rule, where the points clearly separate into two distinct groups. Specifically, for high intraclass correlation (0.8) the prior prediction rule performs extremely poorly. While for lower intraclass correlations, the performance of the prior prediction rule is much closer to that of the multilevel and OLS prediction rules, as expected from Proposition 1.3.

Notes

[1]They examined A/AS level results obtained by English institutions from year to year. Their approach is different from our approach because they are considering cohort periods and are not predicting y_{*j}.

[2]Viewing Equation 1 as a model that describes a hypothetical sequence of replications that generated the data, the introduction of random coefficients expresses the idea that the intercepts and slopes are no longer fixed numbers—which are constant within schools and possibly between schools—and that they may vary over replications (de Leeuw & Kreft, 1995).

[3]Thus, given an estimate $\hat{\gamma}$, the prior estimate for β_j would be $\mathbf{W}_j\hat{\gamma}$.

[4]For an excellent review of estimation of fixed and random effects in the general mixed model, see Robinson, 1991.

[5]See Lindley and Smith (1972) as well for a general treatment of estimation theory in hierarchical models.

[6]Recall that because the Level 1 coefficient β_j is a random variable, the term "estimation" is being employed somewhat pejoratively here.

[7]The shrinkage estimator in Equation 9 is often referred to as a Bayes or posterior estimator.

[8]Recall that we have $\mathbf{y}_j$ and β_j distributed multivariate normal with $E(\mathbf{y}_j) = \mathbf{X}_j\mathbf{W}_j\gamma$, $E(\beta_j) = \mathbf{W}_j\gamma$ and $\text{Cov}(\beta_j, \mathbf{y}_j) = \text{Cov}(\beta_j, \mathbf{X}_j\beta_j + \mathbf{r}_j) = \text{Cov}(\beta_j, \mathbf{X}_j\beta_j) = \tau\mathbf{X}_j'$. And, employing the well known result that the conditional expectation in the normal case is equivalent to the linear regression of β_j on $\mathbf{y}_j$ leads to the result in Equation 11.

[9]Because we are "estimating" a random variable, care must be taken with respect to notation. Given an observed random variable y and an unobservable random variable $\mathbf{w}$, let $t(\mathbf{y})$ be an estimator of the realized value of the random variable w. The MSE of $t(\mathbf{y})$ is defined as $E[t(\mathbf{y}) - \mathbf{w}]^2$, where all expectations are taken with respect to the joint distribution of $\mathbf{y}$ and $\mathbf{w}$. We say that $t(\mathbf{y})$ is unbiased if $E[t(\mathbf{y})] = E(\mathbf{w})$. Given that the prediction error of $t(\mathbf{y})$ equals $t(\mathbf{y}) - \mathbf{w}$, we see that $t(\mathbf{y})$ unbiased implies that the MSE of $t(\mathbf{y})$ equals the variance of its prediction error.

[10]One must restrict oneself to the class of unbiased estimators because a MMSLE does not exist for the unknown γ case (Pfefferman, 1984).

[11]Some authors rebel strongly against the term prediction because the random effects under investigation may have occurred thousands of years ago.

[12]Of course, in the school example neither the student nor the school official is concerned about coefficients estimates, rather, the focus is on the outcome, and the more accurately we can predict the outcome, the better.

[13]Personal communication with D. Rubin, 7/97.

[14]In Rubin's Law School research, the law school officials would be concerned about predictions that are translated too far in the positive direction, while the applicants would be worried about their individual predictions being translated too far in the negative direction.

[15]The MLEs of the Level 1 variance do not exhibit the severe negative bias exhibited by the MLEs of the Level 2 variance.

[16]In spite of the gains in reducing bias, they also comment that the methods exhibited diminished gains for small sample sizes and for extreme values of variance parameters.

[17]Although this result is readily apparent in Morris (1995, p. 193), his different definition of the shrinkage factor causes the shrinkage factor to be equal to $1 - \rho$.

[18]We have a slight abuse of notation because W_j is now a scalar; previously the Level 2 equation in matrix form had $\beta_j = W_j\gamma$, where W_j was a matrix.

[19]The formation of predictive intervals was also employed where we examined the percent of correct coverage over the replications. However, because of the discrete nature of coverage—in the interval or outside the interval—this proved to be less insightful than the continuous measure of predictive mean square error.

[20]The results and graphs from the entire range of simulations are available from the authors upon request.

[21]XLISP-STAT was developed by Luke Tierney and is written in the Xlisp dialect of Lisp, which was developed by David Betz.

[22]An XLISP-STAT program written by James Hilden-Minton, which incorporates both the EM algorithm and Fisher scoring. As noted by Hilden-Minton, while the latter approach is faster, the EM algorithm exhibits greater stability. Initial estimates are obtained from the first iteration of the EM algorithm, after which point the procedure is switched to Fisher scoring and remains with Fisher scoring until convergence unless Fisher scoring produces estimates outside of the parameter space. See "Terrace-Two User's Guide: An XLISP-STAT Package for Estimating Multi-Level Models" by Afshartous and Hilden-Minton (1996) for a full description of Terrace-Two. Software and manuals accessible via World Wide Web site http://www.stat.ucla.edu.

[23]Note that a difference of 0.06 in PMSE is rather large given the range of initial PMSE values, translating to a percentage PMSE reduction of more than 10%.

[24]Note the difference in scale in Figure 3. When originally plotted with the same scale as Figures 1 and 2, the adverse effect of increased J was not as clear. Similar problems existed when Figures 1 and 2 were plotted on the same scale as Figure 3.

[25]The scale for these plots has been chosen to facilitate comparison with the corresponding plots for low intraclass correlation; however, note that for the differential PMSE for the prior prediction rule, Figure 6, we have a different scale because of the much larger PMSE.

References

Afshartous, D. (1995). Determination of sample size for multilevel model design (Tech. Rep. No. 35). In V. S. Williams, L. V. Jones, & I. Olkin (Eds.), *Perspectives on statistics for educational research: Proceedings of the National Institute of Statistical Sciences (NISS).*

Afshartous, D., & de Leeuw, J. (2004). An application of multilevel model prediction to NELS: 88. *Behaviormetrika, 31*(1), 43–66.

Afshartous, D., & de Leeuw, J. (2004). *Decomposition of prediction error in multilevel models.* Manuscript submitted for publication.

Afshartous, D., & de Leeuw, J. (2005). A predictive density approach to predicting future observables in the multilevel model. *Journal of Statistical Planning and Inference, 128,* 149–164.

Afshartous, D., & Hilden-Minton, J. (1996). *TERRACE-TWO: An XLISP-STAT software package for estimating multilevel models: User's guide.* UCLA Department of Statistics Technical Report.

Bassiri, D. (1988). *Large and small sample properties of maximum likelihood estimates for the hierarchical linear model.* Unpublished doctoral dissertation, Department of Counseling, Educational Psychology and Special Education, Michigan State University.

Berger, J. O. (1980). *Statistical decision theory: Foundations, concepts, and methods.* New York: Springer-Verlag.

Browne, W., & Draper, D. (2002). *A comparison of Bayesian and likelihood-based methods for fitting multilevel models.* Retrieved from http://multilevel.ioe.ac.uk/team/bill.html.

Busing, F. (1993). *Distribution characteristics of variance estimates in two-level models* (Tech. Rep. No. PRM 93-04). Leiden, The Netherlands: Department of Psychometrics and Research Methodology, University of Leiden.

Chipman, J. S. (1964). On least squares with insufficient observations. *Journal of the American Statistical Association, 59,* 1078–1111.

de Leeuw, J., & Kreft, I. (1986). Random coefficient models for multilevel analysis. *Journal of Educational Statistics, 11,* 57–86.

de Leeuw, J., & Kreft, I. (1995). Questioning multilevel models. *Journal of the Educational and Behavioral Statistics, 20,* 171–189.

de Leeuw, J., & Kreft, I. (Eds.). (2002). *Handbook of multilevel quantitative analysis.* Boston, Dordrecht, London: Kluwer.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39,* 1–8.

Efron, E., & Morris, C. (1971). Limiting the risk of Bayes and empirical Bayes estimators. Part I: The Bayes case. *Journal of the American Statistical Association, 66,* 807–815.

Efron, E., & Morris, C. (1972). Limiting the risk of Bayes and empirical Bayes estimators. Part II: The empirical Bayes case. *Journal of the American Statistical Association, 67,* 130–139.

Geisser, S. (1979). A predictive approach to model selection. *Journal of the American Statistical Association, 74,* 153–160.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika, 78,* 45–51.

Gray, J., Goldstein, H., Thomas, S. (2001). Predicting the future: The role of past performance in determining trends in institutional effectiveness at A level. *British Educational Research Journal, 27,* 391–406.

Harville, D. A. (1976). Extension of the Gauss Markov Theorem to include the estimation of random effects. *Annals of Statistics, 4,* 384–396.

Hedeker, D., & Gibbons, R. (1996). MIXOR: A Computer program for mixed-effects ordinal probit and logistic regression analysis. *Computer Methods and Programs in Biomedicine, 49,* 157–176.

Hilden-Minton, J. (1995). *Multilevel diagnostics for mixed and hierarchical linear models.* Unpublished doctoral dissertation, UCLA.

Kim, K.-S. (1990). *Multilevel data analysis: A comparative examination of analytical alternatives.* Unpublished doctoral dissertation, Department of Education, University of California, Los Angeles.

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

Littell, R., Milliken, G., Stroup, W., & Wolfinger, R. (1996). *SAS system for mixed models.* Cary, NC: SAS.

Longford, N. T. (1988). Fisher scoring algorithm for variance component analysis of data with multilevel structure. In R. D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 297–310). Orlando, FL: Academic Press.

Marsden, J., & Hoffman, M. (1974). *Elementary classical analysis.* New York: W. H. Freeman.

Mok, M. (1995, June). Sample sizes for 2-level designs in educational research. *Multilevel Modeling Newsletter.*

Morris, C. N. (1995). Hierarchical models for educational data: An overview. *Journal of Educational and Behavioral Statistics, 20,* 190–200.

Morrison, D. F. (1967). *Multivariate statistical methods.* New York: McGraw Hill.

Pfefferman, David. (1984). On extensions of the Gauss–Markov theorem to the case of stochastic regression coefficients. *Journal of the Royal Statistical Society, Series B, 46,* 139–148.

Rao, C. R. (1965). *Linear statistical inference and its applications* (2nd ed.). New York: Wiley.

Rao, C. R. (1987). Prediction of future observations in growth curve models. *Statistical Science, 2,* 434–471.

Rasbash, J., Browne, W., Goldstein, H., Yang, M., et al. (2000). *A user's guide to MLwiN* (2nd ed.). London: Institute of Education.

Raudenbush, S. W. (in press). Many small groups. In J. de Leeuw & I. Kreft (Eds.), *Handbook of quantitative multilevel analysis.* New York: Kluwer.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S. W., Bryk, A. S., Cheong, Y., & Congdon, R. T. (2000). *HLM 5: Hierarchical linear and nonlinear modeling.* Chicago: Scientific Software International.

Robinson, G. K. (1991). That BLUP is a good thing. *Statistical Science, 6,* 15–51.

Rubin, D. (1980). Using empirical Bayes techniques in the Law School Validity Studies. *Journal of the American Statistical Association, 75,* 801–827.

Seltzer, M. (1993). Sensitivity analysis for fixed effects in the hierarchical model: A Gibbs sampling approach. *Journal of Educational and Behavioral Statistics, 18*(3), 207–235.

Spiegelhalter, D. J., Thomas, A., & Best, N. G. (1996). Computation on Bayesian graphical models. *Bayesian Statistics, 5,* 407–425.

Strang, G. (1988). *Linear algebra and its applications.* San Diego: Harcourt Brace Jovanovich.

Swamy, P. A. V. B. (1971). *Statistical inference in a random coefficient model.* New York: Springer.

Authors

DAVID AFSHARTOUS is Assistant Professor, School of Business Administration, University of Miami, Coral Gables, FL 33124-82371; afshar@miami.edu. His areas of interest and specialization are multilevel models, spatial statistics, and wireless communications.

JAN DE LEEUW is Distinguished Professor, and Chair, Department of Statistics, University of California, Los Angeles, CA 90095-1554. His areas of interests and specialization are data analysis, multivariate analysis, and computational statistics.

Manuscript received November 13, 2002
Revision received July 11, 2003
Accepted October 30, 2003