Decomposition of Prediction Error in Multilevel Models[1]

David Afshartous
School of Business Administration, University of Miami,
Coral Gables, FL 33124-8237
Jan de Leeuw
Department of Statistics, University of California,
Los Angeles, CA 90095-1554

ABSTRACT: We present a decomposition of prediction error for the multilevel model in the context of predicting a future observable y_{*j} in the jth group of a hierarchical dataset. The multilevel prediction rule is used for prediction and the components of prediction error are estimated via a simulation study that spans the various combinations of level-1 (individual) and level-2 (group) sample sizes and different intraclass correlation values. These components of prediction error provide information with respect to the cost of parameter estimation versus data imputation for predicting future values in a hierarchical data set. Specifically, the cost of parameter estimation is very small compared to data imputation.

KEY WORDS: prediction error components, Monte Carlo, multilevel model

1 Introduction

Consider the problem of predicting a future observable y_{*j} in the jth group of a hierarchical data set. Various prediction rules may be employed to produce predictions for this future observable. For example, given covariates at the individual or group level, one may cast the problem within the framework of the multilevel model and produce predictions based on OLS, prior, or multilevel approaches. In an earlier study, the performance of these prediction rules was assessed via a large scale simulation study (Afshartous & de Leeuw, 2002). The prediction rule employing a shrinkage estimator proved to be the most accurate. However, this study did not provide any assessment of the components of prediction error for estimating the various parameters that are employed by these prediction rules.

Harville (1985) presented such a decomposition of prediction error framework for the case of the general linear model. We extend this framework to the multilevel model to assess the cost of parameter estimation; in addition, we also consider the cost of data imputation at both the individual and group level. In other words, we are interested in the following two questions: 1) how is our ability to predict y_{*j} affected by the estimation of the model parameters, and 2) how is our ability to predict y_{*j} affected by missing data at either the group level or individual level? Hill & Goldstein (1998) examined the handling of educational data with students belonging to multiple groups and also the case where group membership itself is unknown.[2] Although slightly similar to Hill & Goldstein's problem,

[1]This research was supported by a grant from the National Institute for Statistical Sciences

[2]His method involved developing a cross-classified multilevel model with weights reflecting probabilities of group membership.

our problem centers around unknown information with respect to parameters, individual and group covariates, and its affect on prediction.

In section 1.1 and 1.2 we review the notation and results of the multilevel model. In section 2 we present the decomposition of prediction error framework for the case of the multilevel model. In section 3 we describe the simulation study for estimating the various components of prediction error. Finally, in section 4 we discuss the main results, and in section 5 we present a brief summary and directions for future research.

1.1 The Multilevel Model

Multilevel modeling is a statistical technique designed to facilitate inferences from hierarchical data. A given data point y_{ij} represents the ith case in the jth unit, e.g., the ith student in the jth school for educational data. The multilevel model prediction problem—in its simplest form—consists of predicting a future observable y_{*j}, i.e., a future case of the jth group. For a full review of the multilevel model see Raudenbush & Bryk (2002). We shall restrict this discussion to the simple case of primary units grouped within secondary units and periodically refer to the applied example of students (level-1) grouped within schools (level-2). For example, we may have J schools, where the jth school contains n_j students. The basic multilevel model has the following level-1 model equation:

$$Y_j = X_j\beta_j + r_j, \tag{1}$$

Each X_j has dimensions $n_j \times p$, and $r_j \sim N(0, \sigma^2\Psi_j)$, with Ψ_j usually taken as I_{n_j}. In multilevel modeling, some or all of the level-1 coefficients, β_j, are random variables, and may also be functions of level-2 (school) variables:

$$\beta_j = W_j\gamma + u_j, \tag{2}$$

Each W_j has dimension $p \times q$ and is a matrix of background variables on the jth group, and $u_j \sim N(0, \tau)$. Clearly, since 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.

Combining equations (1) and (2) yields the single equation model:

$$Y_j = X_jW_j\gamma + X_ju_j + r_j \tag{3}$$

which may be viewed as a special case of the mixed linear model, with fixed effects γ and random effects u_j.[3] Thus, marginally, y_j has expected value $X_jW_j\gamma$ and dispersion $V_j = X_j\tau X_j{}' + \sigma^2 I$. 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

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

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

where $d_j = Y_j - X_j W_j \gamma$. Since the J units are independent, we write the log-likelihood for the entire model as a sum of unit log-likelihoods, *i.e.*,

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

1.2 Estimation

Raudenbush & 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. Other approaches focus on the James-Stein "borrowing-of-strength" aspect of multilevel modeling when presenting estimates of the level-1 coefficients.[4]

Another alternative is to focus on the likelihood established by equation 5, where full or restricted 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 $\hat{\beta}_j = (X'_j X_j)^{-1} X_j y_j$ and—given an estimate of γ—the prior estimate $W_j \hat{\gamma}$ 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 which 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 follows:

$$\hat{\beta}_j^{\,*} = \Theta_j \hat{\beta}_j + (I - \Theta_j) W_j \gamma \tag{6}$$

where

$$\Theta_j = \tau(\tau + \sigma^2 (X_j{}' X_j)^{-1})^{-1} \tag{7}$$

is the ratio of the parameter variance for β_j (τ) relative to the variance $\sigma^2 (X_j{}' X_j)^{-1}$ for the OLS estimator for β_j plus this parameter variance matrix. Thus, if the OLS estimate is unreliable, $\hat{\beta}_j^{\,*}$ will pull $\hat{\beta}_j$ towards $W_j \hat{\gamma}$, the prior estimate.[5] Indeed, a little bit of algebra demonstrates that the shrinkage estimator in equation 6 is the expected value of β_j given y_j:[6]

$$\begin{aligned} E(\beta_j | y_j) &= E(\beta_j) + \mathrm{Cov}(\beta_j, y_j)(\mathrm{Var}(y_j))^{-1}(y_j - E(y_j)) \\ &= W_j \gamma + \tau X_j{}' V_j{}^{-1}(y_j - X_j W_j \gamma) \\ &= W_j \gamma + \tau X_j{}' V_j{}^{-1} y_j - \tau X_j{}' V_j{}^{-1} X_j W_j \gamma \end{aligned} \tag{8}$$

[4]Recall that since the level-1 coefficient β_j is a random variable, the term "estimation" is being employed somewhat pejoratively here.

[5]The shrinkage estimator in equation 6 is often referred to as a Bayes or posterior estimator.

[6]Recall that we have y_j and β_j distributed multivariate normal with $E(y_j) = X_j W_j \gamma$, $E(\beta_j) = W_j \gamma$ and $\mathrm{Cov}(\beta_j, y_j) = \mathrm{Cov}(\beta_j, X_j \beta_j + r_j) = \mathrm{Cov}(\beta_j, X_j \beta_j) = \tau X_j{}'$. And, employing the well known result that the conditional expectation in the normal case is equivalent to the linear regression of β_j on y_j leads to the result in equation 8.

Swamy (1971, p.101) presents the following formula for the inverse of V_j,

$$V_j^{-1} = \sigma^2[X_j({X_j}'X_j)^{-1}{X_j}'] + X_j({X_j}'X_j)^{-1}{A_j}^{-1}({X_j}'X_j)^{-1}{X_j}' \tag{9}$$

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

$$\begin{aligned} E(\beta_j|y_j) &= W_j\gamma + \tau{A_j}^{-1}\hat{\beta}_j - \tau{A_j}^{-1}W_j\gamma \\ &= \tau{A_j}^{-1}\hat{\beta}_j + (I - \tau{A_j}^{-1})W_j\gamma \\ &= \Theta_j\hat{\beta}_j + (I - \Theta_j)W_j\gamma \end{aligned}$$

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 1965b).[7]

One may also write the multilevel estimate as $\hat{\beta}_j^* = W_j\gamma + \hat{u}_j$, 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:

$$\hat{u}_j = {C_j}^{-1}{X_j}'(y_j - X_jW_j\gamma) \tag{10}$$

where

$$C_j = {X_j}'X_j + \sigma^2\tau \tag{11}$$

To be sure, 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:

$$\hat{\gamma} = (\sum_{j=1}^{J} {W_j}'{X_j}'{V_j}^{-1}X_jW_j)^{-1}\sum_{j=1}^{J} {W_j}'{X_j}'{V_j}^{-1}y_j \tag{12}$$

where

$$V_j = \text{Var}(y_j) = X_j\tau{X_j}' + \sigma^2 I$$

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 6 employing this estimator of γ yields the minimum mean square linear unbiased estimator (MMSLUE) of β_j (Harville

[7] Since 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 w, let $t(y)$ be an estimator of the realized value of the random variable w. The MSE of $t(y)$ is defined as $E(t(y) - w)^2$, where all expectations are taken with respect to the joint distribution of y and w. We say that $t(y)$ is unbiased if $E(t(y)) = E(w)$. Given that the prediction error of $t(y)$ equals $t(y) - w$, we have that $t(y)$ unbiased implies that the MSE of $t(y)$ equals the variance of its prediction error.

1976).[8] de Leeuw & Kreft (1995) discuss alternative estimates of the fixed effects via a two-step procedure, where one first obtains the OLS estimates of the β_j and then regresses these values on the W_j values. Regardless, this approach of focusing on the estimation of γ instead of β_j is preferred by some since we are actually estimating a parameter and do not may blur 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 prior 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, via procedures such as Fisher Scoring (Longford, 1987), iteratively reweighted generalized least squares (Goldstein, 1986), or the EM algorithm (Dempster, Laird, & Rubin, 1977).[9]

1.3 Multilevel Prediction

Formally, let y_{*j} be the unknown outcome measure for an unsampled observation in the jth group, where group j is not necessarily in our sample or even known. The basic problem as before is to predict y_{*j}. In Afshartous et al. (2002) we assessed the relative performance of three prediction rules in predicting a future observable y_{*j}. The multilevel prediction rule employed the shrinkage estimator above for the level-1 coefficients β_j. Formally:

$$\hat{y}_{*j} = X_{*j}\hat{\beta}_j^{\,*} \tag{13}$$

Note that one may write the multilevel estimate as $\hat{\beta}_j^{\,*} = W_j\hat{\gamma} + \hat{u}_j$, where we recall that u_j may be interpreted in the mixed model sense as the random effect of the jth group. With respect to the prediction of y_{*j}, the predicted value of y_{*j} is $X_{*j}\hat{\beta}_j^{\,*}$, which may also be written as $\hat{y}_{*j} = X_{*j}W_j\hat{\gamma} + X_{*j}\hat{u}_j$. Taking this one step further, we note that Harville (1976) showed that this may also be written as follows:

$$\hat{y}_{*j} = X_{*j}W_j\hat{\gamma} + \hat{V}_{*j}\hat{V}_j^{-1}(y_j - X_jW_j\hat{\gamma}) \tag{14}$$

where

$$\begin{aligned}
\hat{\gamma} &= (\sum_{j=1}^{J} W_j'X_j'\hat{V}_j^{-1}X_jW_j)^{-1}\sum_{j=1}^{J} W_j'X_j'\hat{V}_j^{-1}y_j \\
\hat{V}_{*j} &= \hat{\mathrm{Cov}}(y_{*j}, y_j) = X_{*j}\hat{\tau}X_j' \\
\hat{V}_j &= \hat{\mathrm{Var}}(y_j) = X_j\hat{\tau}X_j' + \hat{\sigma}^2 I
\end{aligned}$$

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

[9] These and other procedures manifest themselves in several software packages: HLM (Raudenbush et al., 2000), MIXOR (Hedeker & Gibbons, 1996), MLWIN (Rabash et al., 2000), SAS Proc Mixed (Littell et al., 1996), and VARCL (Longford, 1988). In addition, the software package BUGS (Spriegelhalter et al., 1994) incorporates fully Bayesian methods that have been introduced (Gelfand et al., 1990; Seltzer, 1993). Note, although Lindley & 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.

This representation illustrates the multilevel prediction rule as the conditional expectation of y_{*j} given the data y.[10] In our previous study, the multilevel prediction rule outperformed prediction rules based on OLS and prior estimators of the level-1 coefficients. Moreover, these results were robust over a very wide simulation design that extensively covered the parameter and sample size space at both level-1 and level-2.

In this paper we take extend these results by applying a decomposition of prediction error framework for the multilevel prediction rule; this extends the results of Harville (1985) for the general linear model. This framework is described in the next section.

2 Decomposition of Multilevel Prediction Error

The questions regarding levels of information with respect to both parameters and data that were discussed earlier are now examined: 1) how is our ability to predict y_{*j} affected by the estimation of the model parameters, and 2) how is our ability to predict y_{*j} affected by missing data at either the group level or individual level? In essence, with respect to the data, the answers to these questions will provide information regarding the relative worth of data at the individual and group level, in addition to the relative costs of estimating the model parameters. We adopt the framework of Harville (1985) in order to examine these questions.

Harville (1985) considered the general problem of predicting of a scalar random variable w from a vector random variable y. Information state 1 is defined as the case where the joint distribution of w and y is known, whereupon the predictor of w is taken as $\mathrm{E}(w|y)$, which has minimum MSE among all predictors. In information state 2, where the first and second moments are known but the joint distribution is unknown, the predictors of w is taken as the linear regression of w on y, which would equal $\mathrm{E}(w|y)$ if the distribution were normal. Harville goes on to develop more predictors of w for additional information states. For example, in information state 3 the second moments are known and the first moments are unknown, and in information state 4 both the first and second moments are unknown.

Below, these states of information are delineated for the the multilevel prediction rule. In addition, the case of "unknown" or missing data is introduced to this framework.[11] For higher information states, unless otherwise specified, parameter estimates are the same as in the previous lower information state.

<u>Info State 2: First and Second Moments Known</u>

$$\hat{y}_{*j} = X_{*j}W_j\gamma + V_{*j}V_j^{-1}(y_j - X_jW_j\gamma) \tag{15}$$

This corresponds to the ideal case where all the necessary parameters are known. As noted earlier, in the normal case one can view this as a conditional expectation. Thus, the parameters that are required by the multilevel prediction rule are known and estimation is

[10] Furthermore, for the case of known γ and known variance components, Rao (1973) showed that $\hat{y}_{*j}$ has minimum MSE among all linear predictors. When γ is estimated as in equation 14 with known variance components, $\hat{y}_{*j}$ has minimum MSE among all linear unbiased predictors, i.e., it is the best linear unbiased predictor (BLUP) (Goldberger, 1962).

[11] Note that for the multilevel model we have information state 1 and information state 2 identical due to the normality assumption, thus we skip information state 1.

unnecessary. For the ensuing simulation study, these parameters are specified beforehand and thus may indeed be substituted into the multilevel prediction rule.

Info State 3: Only Second Moments Known

$$\hat{y}_{*j} = X_{*j}W_j\hat{\gamma} + V_{*j}V_j^{-1}(y_j - X_jW_j\hat{\gamma}) \tag{16}$$

where

$$\hat{\gamma} = (\sum_{j=1}^{J} W_j'X_j'V_j^{-1}X_jW_j)^{-1}\sum_{j=1}^{J} W_j'X_j'V_j^{-1}y_j$$

Here, the coefficient γ must be estimated. However, the estimate should be close to the actual value since the matrices V_{*j} and V_j^{-1} are known. The difference between the performance of the multilevel prediction rule between Info State 2 and Info State 3 may be viewed as an indicator of how well γ is estimated.

Info State 4: First and Second Moments Unknown

$$\hat{y}_{*j} = X_{*j}W_j\hat{\gamma} + \hat{V}_{*j}\hat{V}_j^{-1}(y_j - X_jW_j\hat{\gamma}) \tag{17}$$

where

$$\begin{aligned}
\hat{\gamma} &= (\sum_{j=1}^{J} W_j'X_j'\hat{V}_j^{-1}X_jW_j)^{-1}\sum_{j=1}^{J} W_j'X_j'\hat{V}_j^{-1}y_j \\
\hat{V}_{*j} &= \hat{\text{Cov}}(y_{*j}, y_j) = X_{*j}\hat{\tau}X_j' \\
\hat{V}_j &= \hat{\text{Var}}(y_j) = X_j\hat{\tau}X_j' + \hat{\sigma}^2 I
\end{aligned}$$

The above corresponds to the situation encountered in practice, i.e., all of the model parameters must be estimated from the observed data. The difference between the performance of the multilevel prediction rule between Info State 3 and Info State 4 may be viewed as an indicator of how well the variance components are estimated.

Info State 5: W_j Unknown

$$\hat{y}_{*j} = X_{*j}\overline{W}\hat{\gamma} + \hat{V}_{*j}\hat{V}_j^{-1}(y_j - X_j\overline{W}\hat{\gamma}) \tag{18}$$

where

$$\overline{W} = \frac{1}{J}\sum_{j=1}^{J} W_j$$

With respect to the school example, this corresponds to having missing school data for the student whose outcome variable we wish to predict. We "estimate" or impute this data with the average of the level-2 variables for all the groups. The change in the performance of the multilevel predictor between Info State 4 and Info State 5 is an indicator of how well this missing data is imputed.

Info State 6: X_{*j} Unknown

$$\hat{y}_{*j} = \overline{X}_j W_j \hat{\gamma} + \hat{V}_{*j} \hat{V}_j^{-1} (y_j - X_j \overline{W} \hat{\gamma}) \tag{19}$$

where

$$\begin{aligned} \overline{X}_j &= (1, \frac{1}{n} \sum_{i=1}^{n} X_{ij}) \\ \hat{V}_{*j} &= \overline{X}_j \hat{\tau} X_j' \end{aligned}$$

With respect to our school example, this would correspond to having missing student data for the student whose outcome we wish to predict. We "estimate" or impute this data as the average of the observations in that particular group. The change in the performance of the multilevel predictor between Info State 4 and Info State 6 is an indicator of how well this missing data is imputed.

Each of the information states presented above may be viewed with respect to penalties for missing information. For instance, the difference in prediction between Info State 2 and Info State 3 may can be viewed as the cost of estimating γ. The difference in prediction between Info State 3 and Info State 4 may be viewed in terms of the cost of estimating the variance components. Furthermore, two additional information states have been added to those considered by Harville. How valuable is the level-1 or student level data with respect to prediction? How valuable is the level-2 or school level data with respect to prediction? Insight into these questions may be obtained by examining the performance of the multilevel prediction rule in Info State 5 and 6. To be sure, in all of the above cases, the cost will be underestimated since the correct model is estimated in all cases and thus we have not accounted for how much worse the prediction would have been if our model had been misspecified.[12] By viewing these five information states as yielding five multilevel prediction rules, we investigate the performance of these prediction rules via a simulation study in order to study the aforementioned penalties. The design of the simulation study continues that of our previous design (Afshartous & de Leeuw, 2002), except for the fact that the $J = 300$ and $n = 100$ conditions have been omitted.[13] Note, where before three different prediction rules were compared, now five variations of the same prediction rule are compared. The presentation of results is divided into two sections, one for parameters (Info States 2,3,4) and one for data (Info States 5,6).

3 Simulation Study Design

We measure the components of prediction error via a large scale simulation study. Multilevel data is simulated under a variety of design conditions, closely following the simulation

[12]This model uncertainty issue will be examined in the sequel.

[13]These conditions proved to be unnecessary in the initial simulations, in addition to severely increasing the required computer time.

study of Busing (1993) where the distribution of level-2 variance component estimates was examined. As in Busing (1993), 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:

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + \gamma_{01} W_j + u_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11} W_j + u_{1j}\end{aligned}$$

The γ's are the fixed effects and are set to a predetermined value; they are set all equal to one as in Busing (1993). 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:

$$r_{u_{0j},u_{1j}} = \frac{\tau_{01}}{(\tau_{00}\tau_{11})^{1/2}} \tag{20}$$

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

$$\rho = \frac{\tau_{00}}{\tau_{00} + \sigma^2} \tag{21}$$

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 in order 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:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + \epsilon_{ij} \tag{22}$$

The level-2 outcome variables, the β's, 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 .5. Since only the balanced data case is considered, where there are n units grouped within J groups, a total of Jn outcomes are simulated. In order 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.

Simulations are also 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. To be sure, 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 including an additional group in one's study. We take n ranging from 5 to 100 and J ranging from 10 to 100. For a full description of the entire simulation design, see Appendix A.

Each design specification depends on the level of the parameters and the $J \times n$ sample sizes. There are twenty possible $J \times n$ combinations and twelve possible parameter specifications, yielding a total of 240 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 dataset. In addition, for each design condition 100 replications are performed, i.e., 100 multilevel data sets are simulated for each design condition and prediction is assessed within each of these replications. Thus, since there are 240 design conditions, a total of $24,000$ multilevel data sets will be generated.

This next phase of this simulation study represents a comparison of the components of prediction error mentioned earlier. Recall that the goal is to predict a future observable y_{*j} in each of our J groups and replicate this process 100 times to account for variability. The adequacy of prediction is measured via predictive 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. Thus, for each of the 240 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 and thus the PMSE is the average of 10 squared difference, while for $J = 100$ each replication consists of predicting 100 future observables and thus the PMSE is the average of 100 squared differences. To be sure, since 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.

4 Results

4.1 Parametric Results

The tables in Appendix B include the results for the performance of the multilevel prediction rule under Info States 2,3, and 4 under all design conditions. Aside from when $n = 5$, however, the prediction rules produce average PMSEs that agree to the second decimal place in almost all the design conditions, i.e., there is little penalty for the estimation of the fixed effects and variance components when the group size is 10 or greater. Thus, only the $n = 5$ case is examined in isolation. Table 1 presents the results for the $n = 5$ for various levels of J across the twelve parametric design conditions.

J	n=5
	State 2, State 3, State 4
10	0.3757, 0.3873, 0.4133
25	0.3812, 0.3851, 0.3952
50	0.3833, 0.3853, 0.3897
100	0.3814, 0.3793, 0.3819

Table 1: Mean MSE for Info States 2,3,4

Table 1 clearly indicates the gradual increase in PMSE for the multilevel prediction rule as the information state changes from Info State 2 to Info State 4. This results hold for all levels of J, where there is a slight increase in PMSE across the information states, the magnitude of which decreases as J increases. Indeed, when $J = 100$ there is no difference between the performance of the multilevel prediction rule under these three information states. In fact, there is even an unexpected decrease in PMSE between Info State 2 to Info State 3. Furthermore, note that in all cases the rise in PMSE is quite small, exhibiting a difference in the first decimal place only when $J = 10$ and $n = 5$. Figures 1 - 2 illustrates these results via side-by-side boxplots. In Figure 1, the results are plotted separately for each level of J such that the cases may be studied in isolation. Note that the greatest relative penalties for the multilevel prediction rule occur in the leftmost figure, where $J = 10$. Figure 2 displays the same plot, but this time plotting all of the boxplots on the same scale, allowing one to more easily compare the rise in PMSE across the various levels of J. In addition, this last plot clearly displays the narrow variability of the multilevel prediction rule under each of these information states. Indeed, the variability would be even less for the higher values of group size n (See Appendix B).

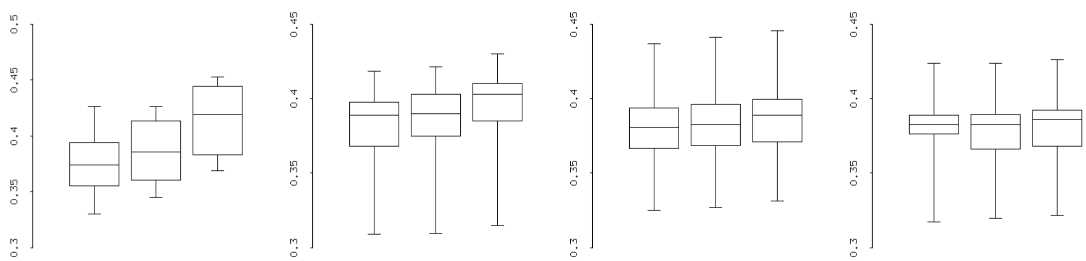

Figure 1: n=5; J=10,25,50,100, MSE for Info States 2,3,4

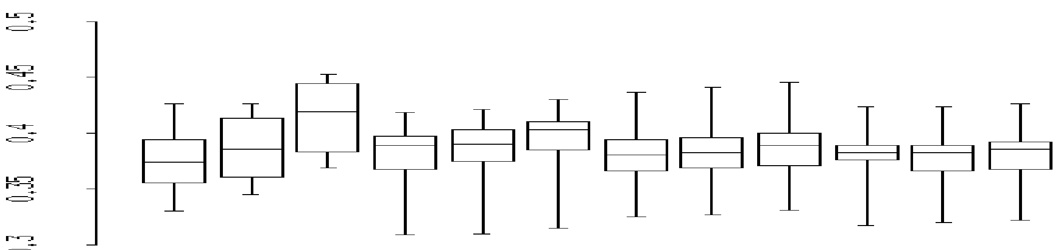

Figure 2: n=5; J=10,25,50,100, MSE for Info States 2,3,4

4.2 Data Results

The results in the previous section indicated little penalties for estimation of unknown parameters. For missing data, however, the results are quite different. The tables in Appendix C display the results for the performance of the multilevel prediction rule under Info States 5 and 6 under all design conditions. In contrast to the results of the previous section, there is a clear increase in the PMSE of the multilevel prediction rule as the information state changes from Info State 4 to Info State 5 and 6. Moreover, this result holds strongly for all levels of $J \times n$ and all twelve parametric design conditions. The results across design conditions for all levels of J and n are shown in Table 2. Info State 4 is included again to enable comparison to the base case which usually exists, i.e., all parameters must be estimated and there is no missing data.

J	n=5	n=10	n=25	n=50
	State 4, State 5, State 6	State 4, State 5, State 6	State 4, State 5, State 6	State 4, State 5, State 6
10	0.4133, 0.7603, 3.4766	0.3159, 0.4790, 3.1769	0.2725, 0.3051, 2.9417	0.2564, 0.2670, 2.9365
25	0.3952, 0.6057, 3.4129	0.3114, 0.3810, 3.1251	0.2757, 0.2810, 3.0541	0.2571, 0.2607, 2.9028
50	0.3897, 0.5714, 3.4079	0.3120, 0.3645, 3.2233	0.2745, 0.2835, 2.9996	0.2620, 0.2642, 2.8855
100	0.3819, 0.5570, 3.4319	0.3091, 0.3562, 3.1530	0.2714, 0.2792, 3.0019	0.2617, 0.2635, 2.9158

Table 2: Mean MSE for Info States 4,5,6

Table 2 indicates that the cost of missing level-1 information is clearly higher than the cost of missing level-2 information for all levels of J and n. This result is more apparent in the side-by-side boxplots presented in Figures 3 - 6. The PMSE produced by the multilevel prediction rule with missing level-2 information (Info State 4) has a distribution similar in level and spread to that produced by the multilevel prediction rule in the base case (Info State 4), while that produced with missing level-1 information (Info State 6) exhibits both a higher level and spread in the boxplots; this last result holds even for large n where one would expect a fairly reliable imputation of the missing level-1 information with the many level-1 observed units.

Table 2 indicates that although there is clearly a large cost for missing level-2 information, this cost decreases monotonically with n for each level of J. The monotonic reduction of the cost of missing information as n rises also holds for missing level-1 information (Info State 6)—as one would expect since the imputation of the missing data relies on more data—albeit the proportional reduction is not as much as that which is exhibited for the missing level-2 information case. From the perspective of data imputation, this is somewhat of a surprise since one would expect the missing level-1 information to be better imputed as n rises, whereas one would expect the missing level-2 imputation to be independent of n since there is only one level-2 observation per group. A possible explanation is the following: as n increases, so does the reliability of our OLS estimate and hence its relative weight with respect to the prior estimate and, since the OLS estimate doesn't involve W_j, this explains the result of the decreased cost of missing W_j as n increases.[14] The effect of increased n on the performance of the multilevel prediction rule under missing level-1 and level-2 information is presented via side-by-side boxplots in Figures 7 - 8. In addition to illustrating the aforementioned results, the boxplots nicely add the information not included in the table: The spread of the PMSE produced by the multilevel prediction rule in the presence of missing level-2 information (Info State 5) decreases as n rises, whereas such is not the case in the case in the presence of missing level-1 information (Info State 6), once again slightly counter-intuitive given the manner in which the missing information has been imputed in both cases.

Table 2 also indicates that for each level of n an increase in J provides a slight reduction

[14]Recall that the multilevel estimate of β_j is a weighted average of the OLS and prior estimate, withe the weights depending to the reliability of the OLS estimate and the prior variance of β_j.

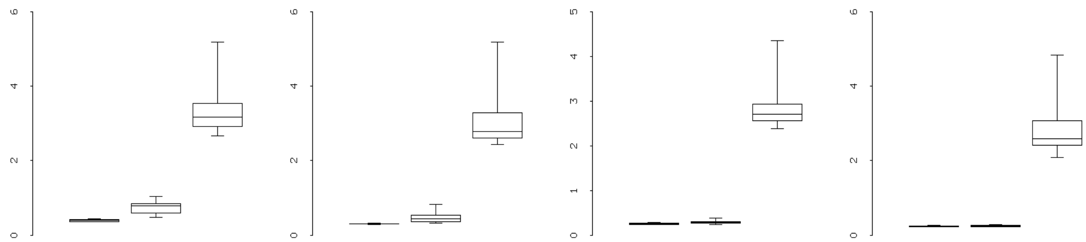

Figure 3: J=10; n=5,10,25,50; MSE for Info States 4,5,6

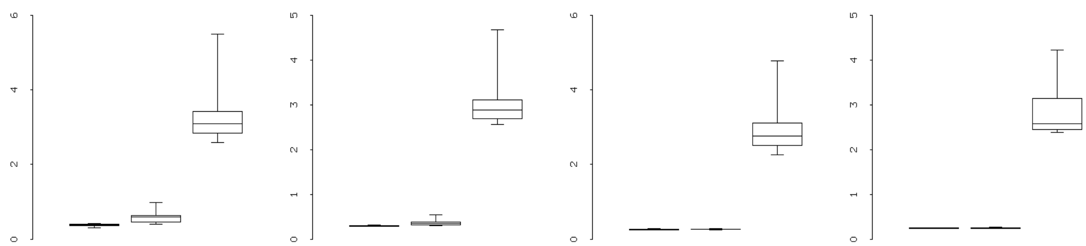

Figure 4: J=25; n=5,10,25,50; MSE for Info States 4,5,6

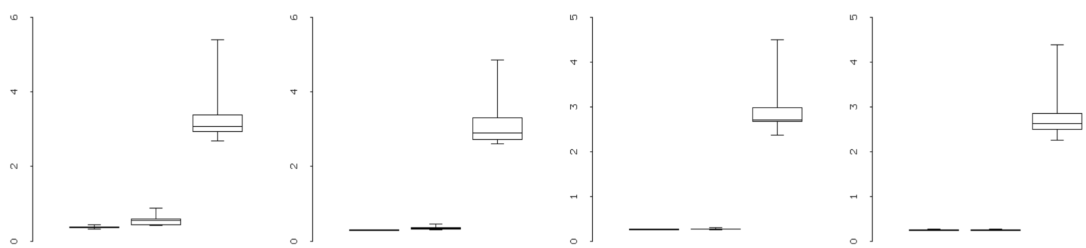

Figure 5: J=50; n=5,10,25,50; MSE for Info States 4,5,6

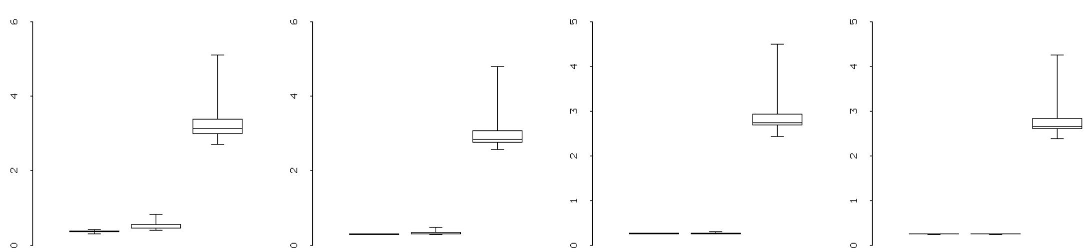

Figure 6: J=100; n=5,10,25,50; MSE for Info States 4,5,6

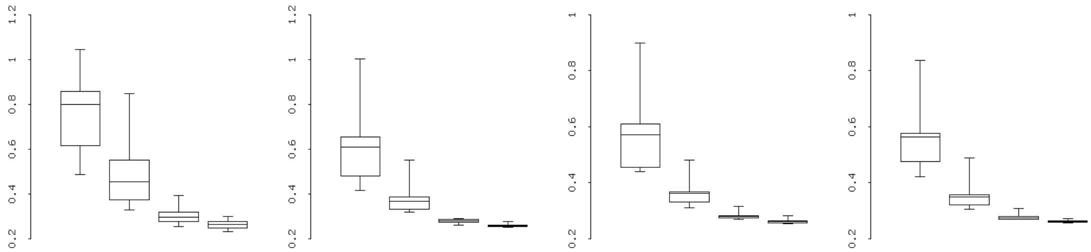

Figure 7: J=10,25,50,100; MSE for Info State 5 as n=5,10,25,50

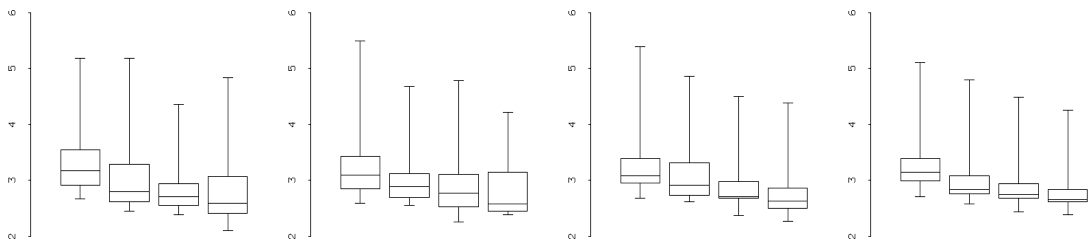

Figure 8: J=10,25,50,100; MSE for Info State 6 as n=5,10,25,50

in the PMSE produced by the multilevel prediction rule with missing level-2 information (Info State 5), although this result is negligible when $n = 50$. For the multilevel prediction rule with missing level-1 information (Info State 6), however, this result does not hold, i.e., for fixed n an increase in J does not produce appreciable reductions in PMSE. The effect of increased J on the performance of the multilevel prediction rule under missing level-1 and level-2 information is presented via side-by-side boxplots in Figures 9 - 10. In addition to illustrating the aforementioned results, the boxplots once again nicely add the information about the spread of the PMSE produced by the multilevel prediction rule with missing level-2 information (Info State 5). While the spread of PMSE produced by the multilevel prediction rule with missing level-2 information is reduced as J increases, such is not the case with missing level-1 information. Furthermore, the boxplots demonstrate that the effect of J in reducing PMSE seems to be less than that of n for both situations.

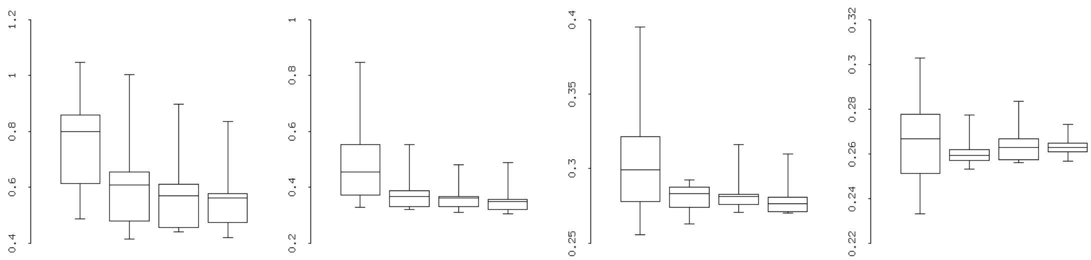

Figure 9: n=5,10,25,50; MSE for info state 5 as J=10,25,50,100

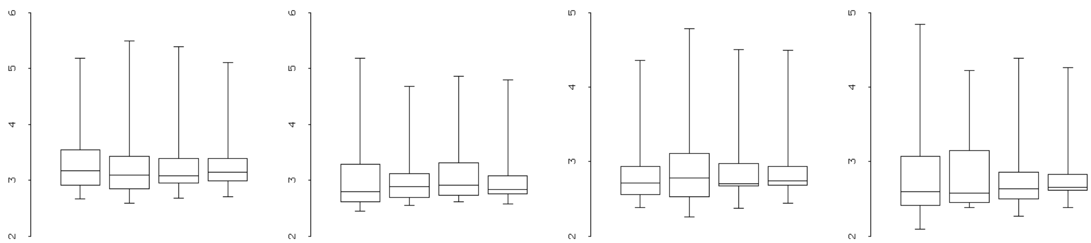

Figure 10: n=5,10,25,50; MSE for info state 6 as J=10,25,50,100

The use of a three-dimensional display provides additional insight into this disparity between the performance of the multilevel prediction rule with missing level-1 and missing level-2 information with respect to the effect of J and n. In the three-dimensional plots of Figure 11, we see the aforementioned results illustrated in three dimensions. For instance, for the missing level-2 data case (Info State 5), the slope of the surface is greater than that for the missing level-2 case (Info State 5) in the direction of increased n. Also, there is little slope in the direction of increased J for the missing level-1 data case while there is a noticeable slope for this direction for the missing level-2 data case. The display for Info State 4 is included for comparative purposes, giving a sense of how the original base situation gets distorted.

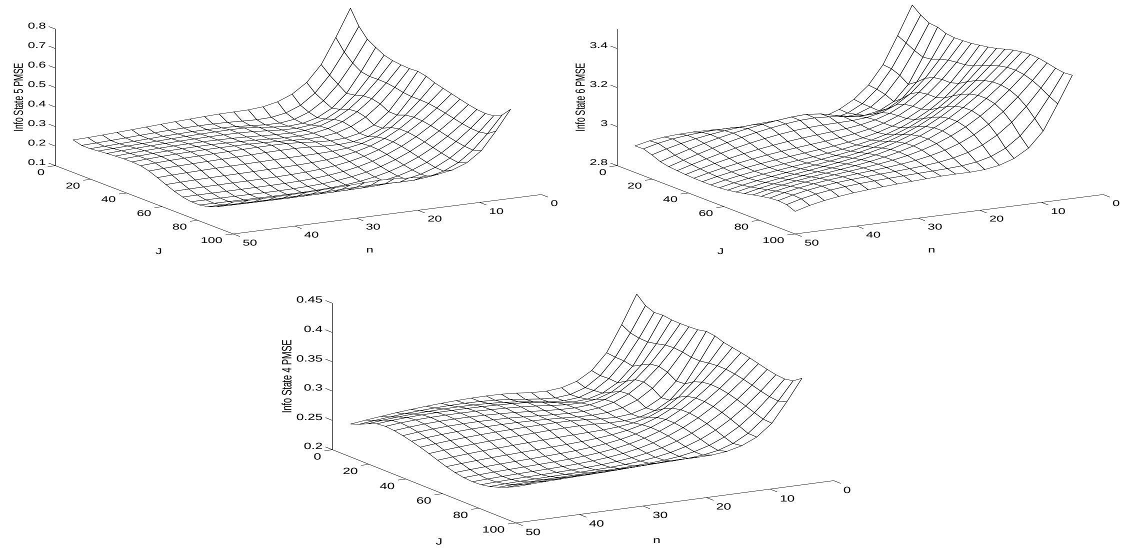

Figure 11: PMSE for Info States 4,5,6

5 Summary

We have presented a decomposition of prediction error for the multilevel model in the context of predicting a future observable y_{*j} in the jth group of a hierarchical dataset. The multilevel prediction rule is used for prediction and the components of prediction error are estimated via a simulation study that spans the various combinations of level level-1 and level-2 sample sizes and different intraclass correlation values. These components of prediction provide information with respect to the cost of parameter estimation versus data imputation for predicting future values in a hierarchical data set. Namely, the cost of parameter estimation is very small compared to data imputation. To be sure, these results are specific to our design space and may vary across other design spaces. Specific results are enumerated below:

1. The performance of multilevel prediction rule exhibits little variation across Info States 2,3, and 4 under all design conditions. The only exception being when group size $n = 5$, where a slight increase in PMSE occurs for estimating both the fixed effects and variance components.

2. The magnitude of the increase in PMSE across the information states 2,3, and 4 decreases as the number of groups J increases.

3. There is a clear increase in the PMSE of the multilevel prediction rule as the information state changes from Info State 4 to Info States 5 and 6 for all design conditions. In other words, data imputation is much more costly than parameter estimation.

4. The cost of missing level-1 information is higher than that of missing level-2 information for all levels of J and n. Thus, with respect to prediction, it is much more costly to have individual level info missing than group level information.

5. The multilevel prediction rule is more responsive to increases in group size n in the presence of missing of level-2 data (Info State 5) than in the presence of level-1 data (Info State 6), a somewhat counter-intuitive result. In other words, if data is missing at the group level, prediction is more improved as n increases as compared to the case when data is missing at the individual level.

6. An increase in the number of groups J provides a slight reduction in the PMSE produced by the multilevel prediction rule with missing level-2 information (Info State 5); The corresponding result does not hold for missing level-1 information (Info State 6). In other words, if data is missing at the group level, prediction improves as the number of groups increases, whereas if data is missing at the individual level this is not the case.[15]

To be sure, in all of the above cases, the costs will be underestimated since the correct model is estimated; thus, we have not accounted for how much worse the prediction would have been if our model had been mis-specified. This model uncertainty issue will be examined in future research, where we investigate the impact of mis-specifying both the level-1 and level-2 model equation.

[15] To be sure, this and the previous result are affected by the manner in which the missing data is imputed.

A Simulation Design

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

Table 3: ρ, τ_{00}, τ_{11}

	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

Table 4: τ_{01}

	n_j				
J	5	10	25	50	100
10	50	100	250	500	1000
25	125	250	625	1250	2500
50	250	500	1250	2500	5000
100	500	1000	2500	5000	10000

Table 5: Sample sizes

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

Table 6: Design numbers

B PE Decomposition: Info States 2,3,4

J	n=5	n=10	n=25	n=50
	State 2, State 3, State 4	State 2, State 3, State 4	State 2, State 3, State 4	State 2, State 3, State 4
10	0.3562, 0.3733, 0.3981	0.2885, 0.2909, 0.3004	0.2415, 0.2424, 0.2461	0.2703, 0.2708, 0.2718
25	0.3579, 0.3598, 0.3743	0.2878, 0.2916, 0.2952	0.2490, 0.2489, 0.2492	0.2490, 0.2489, 0.2492
50	0.3582, 0.3616, 0.3636	0.2962, 0.2971, 0.2989	0.2778, 0.2779, 0.2784	0.2585, 0.2587, 0.2589
100	0.3481, 0.3485, 0.3505	0.3114, 0.3120, 0.3131	0.2746, 0.2745, 0.2749	0.2655, 0.2655, 0.2657

Table 7: Design #1: Mean MSE for 2,3,4 Info States

J	n=5	n=10	n=25	n=50
	State 2, State 3, State 4	State 2, State 3, State 4	State 2, State 3, State 4	State 2, State 3, State 4
10	0.3554, 0.3541, 0.3708	0.3136, 0.3128, 0.3301	0.2650, 0.2644, 0.2662	0.2819, 0.2817, 0.2822
25	0.4031, 0.4081, 0.4231	0.3115, 0.3119, 0.3140	0.2639, 0.2639, 0.2644	0.2605, 0.2605, 0.2609
50	0.3969, 0.4026, 0.4061	0.2994, 0.3000, 0.3022	0.2752, 0.2751, 0.2753	0.2634, 0.2634, 0.2633
100	0.3935, 0.3943, 0.3975	0.3076, 0.3080, 0.3088	0.2658, 0.2658, 0.2657	0.2581, 0.2581, 0.2582

Table 8: Design #2: Mean MSE for 2,3,4 Info States

J	n=5	n=10	n=25	n=50
	State 2, State 3, State 4	State 2, State 3, State 4	State 2, State 3, State 4	State 2, State 3, State 4
10	0.3414, 0.3580, 0.3899	0.3098, 0.3107, 0.3249	0.2749, 0.2748, 0.2753	0.2630, 0.2635, 0.2637
25	0.3957, 0.3935, 0.4071	0.3253, 0.3238, 0.3273	0.2664, 0.2659, 0.2665	0.2543, 0.2544, 0.2547
50	0.3776, 0.3765, 0.3794	0.3133, 0.3133, 0.3163	0.2769, 0.2767, 0.2775	0.2620, 0.2620, 0.2620
100	0.3828, 0.3841, 0.3876	0.3049, 0.3054, 0.3067	0.2769, 0.2769, 0.2771	0.2621, 0.2622, 0.2622

Table 9: Design #3: Mean MSE for 2,3,4 Info States

J	n=5	n=10	n=25	n=50
	State 2, State 3, State 4	State 2, State 3, State 4	State 2, State 3, State 4	State 2, State 3, State 4
10	0.3830, 0.4069, 0.4245	0.3008, 0.3044, 0.3122	0.2886, 0.2883, 0.2899	0.2334, 0.2336, 0.2332
25	0.4002, 0.4058, 0.4136	0.3121, 0.3125, 0.3171	0.2764, 0.2760, 0.2768	0.2543, 0.2544, 0.2546
50	0.3856, 0.3871, 0.3922	0.3139, 0.3140, 0.3157	0.2755, 0.2754, 0.2758	0.2607, 0.2608, 0.2609
100	0.3776, 0.3794, 0.3821	0.3128, 0.3128, 0.3135	0.2651, 0.2650, 0.2652	0.2555, 0.2556, 0.2555

Table 10: Design #4: Mean MSE for 2,3,4 Info States

J	n=5	n=10	n=25	n=50
	State 2, State 3, State 4	State 2, State 3, State 4	State 2, State 3, State 4	State 2, State 3, State 4
10	0.3929, 0.4103, 0.4427	0.3002, 0.3037, 0.3119	0.2793, 0.2787, 0.2825	0.2377, 0.2381, 0.2384
25	0.3948, 0.4007, 0.4084	0.3032, 0.3029, 0.3093	0.2713, 0.2710, 0.2725	0.2571, 0.2569, 0.2570
50	0.3823, 0.3855, 0.3918	0.3189, 0.3184, 0.3197	0.2743, 0.2742, 0.2746	0.2649, 0.2648, 0.2649
100	0.3846, 0.3848, 0.3871	0.3074, 0.3077, 0.3082	0.2743, 0.2742, 0.2739	0.2610, 0.2609, 0.2610

Table 11: Design #5: Mean MSE for 2,3,4 Info States

J	n=5	n=10	n=25	n=50
	State 2, State 3, State 4	State 2, State 3, State 4	State 2, State 3, State 4	State 2, State 3, State 4
10	0.3719, 0.3894, 0.4189	0.3117, 0.3120, 0.3182	0.2839, 0.2865, 0.2889	0.2550, 0.2547, 0.2560
25	0.3786, 0.3874, 0.4019	0.3002, 0.3005, 0.3049	0.2711, 0.2714, 0.2714	0.2559, 0.2562, 0.2567
50	0.3799, 0.3806, 0.3863	0.3140, 0.3144, 0.3167	0.2722, 0.2724 ,0.2724	0.2552, 0.2551, 0.2551
100	0.3754, 0.3767, 0.3788	0.3137, 0.3137, 0.3143	0.2708, 0.2707, 0.2708	0.2637, 0.2637, 0.2638

Table 12: Design #6: Mean MSE for 2,3,4 Info States

J	n=5	n=10	n=25	n=50
	State 2, State 3, State 4	State 2, State 3, State 4	State 2, State 3, State 4	State 2, State 3, State 4
10	0.3962, 0.4174, 0.4534	0.3050, 0.3089, 0.3181	0.2573, 0.2564, 0.2567	0.2615, 0.2620, 0.2629
25	0.3572, 0.3648, 0.3758	0.2925, 0.2924, 0.2954	0.2846, 0.2846, 0.2851	0.2549, 0.2544, 0.2546
50	0.3755, 0.3779, 0.3826	0.3154, 0.3161, 0.3180	0.2656, 0.2657, 0.2665	0.2563, 0.2563, 0.2564
100	0.3829, 0.3840, 0.3864	0.3109, 0.3110, 0.3119	0.2728, 0.2728, 0.2729	0.2608, 0.2608, 0.2609

Table 13: Design #7: Mean MSE for 2,3,4 Info States

J	n=5	n=10	n=25	n=50
	State 2, State 3, State 4	State 2, State 3, State 4	State 2, State 3, State 4	State 2, State 3, State 4
10	0.4265, 0.4268, 0.4473	0.3136, 0.3146, 0.3189	0.2731, 0.2733, 0.2734	0.2746, 0.2746, 0.2749
25	0.4189, 0.4216, 0.4306	0.3192, 0.3195, 0.3205	0.2908, 0.2907, 0.2910	0.2535, 0.2536, 0.2535
50	0.4370, 0.4356, 0.4426	0.3084, 0.3083, 0.3088	0.2811, 0.2811, 0.2812	0.2565, 0.2564, 0.2564
100	0.4243, 0.4240, 0.4265	0.3124, 0.3121, 0.3122	0.2714, 0.2714, 0.2714	0.2620, 0.2620, 0.2620

Table 14: Design #8: Mean MSE for 2,3,4 Info States

J	n=5	n=10	n=25	n=50
	State 2, State 3, State 4	State 2, State 3, State 4	State 2, State 3, State 4	State 2, State 3, State 4
10	0.3306, 0.3457, 0.3694	0.2998, 0.3059, 0.3194	0.2560, 0.2557, 0.2590	0.2458, 0.2457, 0.2456
25	0.3092, 0.3097, 0.3151	0.2985, 0.3019, 0.3061	0.2622, 0.2629, 0.2634	0.2673, 0.2673, 0.2683
50	0.3252, 0.3273, 0.3315	0.2975, 0.2982, 0.3007	0.2718, 0.2719, 0.2725	0.2779, 0.2781, 0.2785
100	0.3177, 0.3201, 0.3221	0.2986, 0.2989, 0.2997	0.2670, 0.2673, 0.2680	0.2622, 0.2623, 0.2625

Table 15: Design #9: Mean MSE for 2,3,4 Info States

J	n=5	n=10	n=25	n=50
	State 2, State 3, State 4	State 2, State 3, State 4	State 2, State 3, State 4	State 2, State 3, State 4
10	0.3652, 0.3634, 0.3768	0.2926, 0.2933, 0.3009	0.3013, 0.3000, 0.3030	0.2530, 0.2526, 0.2533
25	0.3804, 0.3860, 0.3951	0.3003, 0.3009, 0.3049	0.2770, 0.2776, 0.2784	0.2580, 0.2576, 0.2583
50	0.3543, 0.3565, 0.3595	0.3085, 0.3085, 0.3108	0.2693, 0.2695, 0.2701	0.2692, 0.2695, 0.2697
100	0.3551, 0.3564, 0.3579	0.3068, 0.3069, 0.3079	0.2679, 0.2678, 0.2682	0.2637, 0.2637, 0.2636

Table 16: Design #10: Mean MSE for 2,3,4 Info States

J	n=5	n=10	n=25	n=50
	State 2, State 3, State 4	State 2, State 3, State 4	State 2, State 3, State 4	State 2, State 3, State 4
10	0.4125, 0.4191, 0.4479	0.2972, 0.2984, 0.3087	0.2719, 0.2727, 0.2745	0.2322, 0.2326, 0.2330
25	0.3843, 0.3868, 0.3966	0.3155, 0.3148, 0.3175	0.2746, 0.2747, 0.2756	0.2600, 0.2600, 0.2600
50	0.3916, 0.3906, 0.3942	0.3107, 0.3103, 0.3105	0.2701, 0.2703, 0.2702	0.2558, 0.2557, 0.2557
100	0.3824, 0.3823, 0.3863	0.3070, 0.3064, 0.3071	0.2730, 0.2729, 0.2728	0.2645, 0.2645, 0.2645

Table 17: Design #11: Mean MSE for 2,3,4 Info States

J	n=5	n=10	n=25	n=50
	State 2, State 3, State 4	State 2, State 3, State 4	State 2, State 3, State 4	State 2, State 3, State 4
10	0.3770, 0.3834, 0.4200	0.3234, 0.3241, 0.3273	0.2534, 0.2532, 0.2546	0.2611, 0.2612, 0.2617
25	0.3942, 0.3973, 0.4049	0.3208, 0.3217, 0.3241	0.2694, 0.2693, 0.2696	0.2575, 0.2574, 0.2574
50	0.4359, 0.4416, 0.4461	0.3236, 0.3235, 0.3252	0.2815, 0.2815, 0.2817	0.2626, 0.2626, 0.2626
100	0.4169, 0.4169, 0.4196	0.3060, 0.3060, 0.3065	0.2764, 0.2765, 0.2765	0.2603, 0.2604, 0.2603

Table 18: Design #12: Mean MSE for 2,3,4 Info States

C PE Decomposition: Info States 5 and 6

J	n=5	n=10	n=25	n=50
	State 5, State 6	State 5, State 6	State 5, State 6	State 5, State 6
10	1.0482, 2.9057	0.8515, 3.0813	0.3955, 2.6252	0.3032, 2.1040
25	1.0054, 2.9095	0.5537, 2.5652	0.2632, 2.3924	0.2632, 2.3924
50	0.8999, 2.7029	0.4821, 2.6432	0.3167, 2.3791	0.2670, 2.2713
100	0.8377, 2.8926	0.4903, 2.6157	0.3100, 2.4805	0.2735, 2.3967

Table 19: Design #1: Mean MSE for 5,6 Info States

J	n=5	n=10	n=25	n=50
	State 5, State 6	State 5, State 6	State 5, State 6	State 5, State 6
10	0.8152, 2.9450	0.4835, 2.7549	0.2856, 2.4425	0.2940, 2.9652
25	0.6186, 3.0928	0.3730, 2.9041	0.2726, 2.7537	0.2641, 2.8561
50	0.6111, 3.0499	0.3393, 2.6320	0.2815, 2.7032	0.2658, 2.5886
100	0.5667, 3.0512	0.3572, 2.9353	0.2715, 2.7411	0.2596, 2.6266

Table 20: Design #2: Mean MSE for 5,6 Info States

J	n=5	n=10	n=25	n=50
	State 5, State 6	State 5, State 6	State 5, State 6	State 5, State 6
10	0.7858, 2.6833	0.5668, 2.9422	0.2932, 2.7693	0.2672, 2.6462
25	0.6040, 3.3108	0.3978, 3.0486	0.2761, 2.8535	0.2569, 2.5755
50	0.5781, 3.0962	0.3627, 2.9070	0.2829, 2.7210	0.2638, 2.6045
100	0.5626, 3.2118	0.3505, 2.8562	0.2839, 2.7679	0.2636, 2.6896

Table 21: Design #3: Mean MSE for 5,6 Info States

J	n=5	n=10	n=25	n=50
	State 5, State 6	State 5, State 6	State 5, State 6	State 5, State 6
10	0.8526, 3.1785	0.4570, 2.5396	0.3123, 2.7564	0.2425, 2.4545
25	0.6771, 3.1804	0.3733, 2.8898	0.2889, 2.8760	0.2571, 2.4261
50	0.5847, 3.0764	0.3632, 2.9765	0.2836, 2.7159	0.2616, 2.7238
100	0.5813, 3.0149	0.3544, 2.7939	0.2704, 2.7171	0.2568, 2.6450

Table 22: Design #4: Mean MSE for 5,6 Info States

J	n=5	n=10	n=25	n=50
	State 5, State 6	State 5, State 6	State 5, State 6	State 5, State 6
10	0.8365, 3.4764	0.4015, 2.6129	0.3060, 2.6831	0.2448, 2.5641
25	0.6373, 3.0595	0.3794, 2.6843	0.2850, 2.8196	0.2584, 2.6241
50	0.6139, 3.0994	0.3698, 2.9366	0.2826, 2.7131	0.2669, 2.7036
100	0.5752, 3.2387	0.3505, 2.7525	0.2786, 2.7424	0.2620, 2.6515

Table 23: Design #5: Mean MSE for 5,6 Info States

J	n=5	n=10	n=25	n=50
	State 5, State 6	State 5, State 6	State 5, State 6	State 5, State 6
10	0.7669, 3.1023	0.5422, 2.7693	0.3228, 2.5093	0.2668, 2.3932
25	0.6273, 2.7814	0.3675, 3.0117	0.2852, 2.4778	0.2606, 2.4933
50	0.5586, 2.8606	0.3693, 3.0082	0.2791, 2.7565	0.2566, 2.4491
100	0.5366, 3.1509	0.3588, 2.8363	0.2768, 2.7731	0.2653, 2.6713

Table 24: Design #6: Mean MSE for 5,6 Info States

J	n=5	n=10	n=25	n=50
	State 5, State 6	State 5, State 6	State 5, State 6	State 5, State 6
10	0.8981, 3.5148	0.4551, 2.6276	0.2754, 2.6399	0.2757, 2.5360
25	0.5796, 2.6011	0.3534, 2.7101	0.2928, 2.5903	0.2585, 2.6046
50	0.5669, 3.0529	0.3636, 2.8203	0.2733, 2.6932	0.2588, 2.6806
100	0.5732, 3.1577	0.3534, 2.8705	0.2783, 2.6753	0.2624, 2.6161

Table 25: Design #7: Mean MSE for 5,6 Info States

J	n=5	n=10	n=25	n=50
	State 5, State 6	State 5, State 6	State 5, State 6	State 5, State 6
10	0.4890, 5.0581	0.3299, 5.1909	0.2755, 4.2158	0.2756, 4.8460
25	0.4494, 5.5024	0.3224, 4.6908	0.2917, 4.7966	0.2535, 4.1796
50	0.4571, 5.0561	0.3112, 4.7566	0.2818, 4.5092	0.2565, 4.1857
100	0.4396, 5.1122	0.3143, 4.7095	0.2714, 4.3030	0.2621, 4.2673

Table 26: Design #8: Mean MSE for 5,6 Info States

J	n=5	n=10	n=25	n=50
	State 5, State 6	State 5, State 6	State 5, State 6	State 5, State 6
10	0.8671, 2.8750	0.5744, 2.4527	0.3366, 2.3921	0.2803, 2.1780
25	0.6995, 2.7996	0.4588, 2.6434	0.2872, 2.2678	0.2777, 2.4261
50	0.6407, 2.6910	0.4225, 2.6848	0.2952, 2.4095	0.2838, 2.4429
100	0.6422, 2.7105	0.3949, 2.5877	0.2879, 2.4460	0.2675, 2.4873

Table 27: Design #9: Mean MSE for 5,6 Info States

J	n=5	n=10	n=25	n=50
	State 5, State 6	State 5, State 6	State 5, State 6	State 5, State 6
10	0.6950, 3.1905	0.3990, 2.8337	0.3206, 2.8883	0.2579, 2.9453
25	0.5136, 3.1256	0.3382, 2.8513	0.2822, 2.6964	0.2605, 2.5714
50	0.4583, 3.1620	0.3432, 2.8083	0.2727, 2.6806	0.2707, 2.5665
100	0.4589, 2.9972	0.3289, 2.8156	0.2717, 2.7593	0.2645, 2.6963

Table 28: Design #10: Mean MSE for 5,6 Info States

J	n=5	n=10	n=25	n=50
	State 5, State 6	State 5, State 6	State 5, State 6	State 5, State 6
10	0.5385, 3.5950	0.3503, 3.504	0.2814, 3.0082	0.2335, 3.2013
25	0.4396, 3.5683	0.3292, 3.2032	0.2778, 3.3571	0.2608, 3.4623
50	0.4425, 3.6504	0.3205, 3.6368	0.2711, 3.2156	0.2564, 3.0163
100	0.4221, 3.5693	0.3137, 3.2499	0.2737, 3.1242	0.2645, 2.9907

Table 29: Design #11: Mean MSE for 5,6 Info States

J	n=5	n=10	n=25	n=50
	State 5, State 6	State 5, State 6	State 5, State 6	State 5, State 6
10	0.5305, 5.1948	0.3400, 4.8137	0.2562, 4.3705	0.2627, 4.4045
25	0.4174, 5.0258	0.3257, 4.2982	0.2698, 4.7673	0.2576, 4.2259
50	0.4452, 5.4005	0.3273, 4.8774	0.2815, 4.4988	0.2625, 4.3936
100	0.4336, 5.0754	0.3070, 4.8129	0.2765, 4.5083	0.2603, 4.2580

Table 30: Design #12: Mean MSE for 5,6 Info States

References

Afshartous, David & de Leeuw, Jan (2002). "Prediction in Multilevel Models," *submitted to refereed journal for review.*

Afshartous, David (1997). *Prediction in Multilevel Models*, unpublished Ph.D. dissertation, UCLA.

Afshartous, David. & Hilden-Minton, James (1996). "TERRACE-TWO: An XLISP-STAT Software Package for Estimating Multilevel Models: User's Guide," *U.C.L.A Department of Statistics Technical Report.*

Atchinson J. (1975). "Goodness of Prediction Fit," *Biometrika*, 62, pp.547-554.

Bryk, A. & Raudenbush, S. (1992). *Hierarchical Linear Models*, Sage Publications, Newbury Park.

Butler, Ronald W. (1986). "Predictive Likelihood with Applications," *Journal of the Royal Statistical Society, Series B*, 48, pp.1-38.

Busing, F. (1993). "Distribution Characteristics of Variance Estimates in Two-level Models," Technical Report PRM 93-04, Department of Psychometrics and Research Methodology, University of Leiden, Leiden, Netherlands.

Chipman, J.S. (1964). "On Least Squares with Insufficient Observations," *Journal of the American Statistical Association* , 59, pp.1078-1111.

de Leeuw, Jan & Kreft, Ita, eds. (2002). *Handbook of Multilevel Quantitative Analysis.* Boston, Dordrecht, London: Kluwer Academic Publishers. *In Press.*

de Leeuw, Jan & Kreft, Ita (1995). "Questioning Multilevel Models," *Journal of the Educational and Behavioral Statistics*, 20, pp.171-189.

de Leeuw, Jan & Kreft, Ita (1986). "Random Coefficient Models for Multilevel Analysis," *Journal of Educational Statistics*, 11, pp.57-86.

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, pp.1-8.

Gelfand, A., & Smith, A. (1990). "Sampling Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, pp.398-409.

Geisser, Seymour (1971). "The Inferential Use of Predictive Distributions," in *Foundations of Statistical Inference*, eds., V.P. Godambe and D.A. Sprott, pp.456-469. Toronto: Holt, Rhinehart, and Winston.

Geisser, S. (1979). "A Predictive Approach to Model Selection," *Journal of the American Statistical Association* , 74, pp.153-160.

Goldberger, A.S. (1962). "Best Linear Unbiased Prediction in the General Linear Model," *Journal of the American Statistical Association*, 57, p.369-375.

Goldstein, H. (1986). "Multilevel Mixed Linear Model Analysis using Iterative Generalized Least Squares," *Biometrika*, 78, pp45-51.

Gotway, C. & Cressie, N. (1993). "Improved Multivariate Prediction under a General Linear Model," *Journal of Multivariate Analysis*, 45, 56-72.

Gray, J., Goldstein, H., Thomas, S. (2001). "Predicting the Future: The Role of Past Performance in Determining Trends in Institutional Effectiveness," *Multilevel Models Project web page, www.ioe.ac.uk/multilevel./.*

Harville, David A. (1985). "Decomposition of Prediction Error," *Journal of the American Statistical Association*, 80, p.132-138.

Harville, David A. (1976). "Extension of the Gauss Markov Theorem to Include the Estimation of Random Effects," *Annals of Statistics*, 4, p.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, pp.157-176.

Hilden-Minton, James (1995). *Multilevel Diagnostics for Mixed and Hierarchical Linear Models*, unpublished Ph.D. dissertation, UCLA.

Hill, Peter W., Goldstein, Harvey. (1998). "Multilevel Modelling of Educational Data with Cross-Classification and Missing Identification for Unites," Journal of Educational and Behavioral Statistics, 23, pp.??

Kullback, S. & Leibler, R.A. (1951). "On Information and Sufficiency," *Annals of Mathematical Statistics*, 22, pp.525-540.

Larimore, W.E. (1983). "Predictive Inference, Sufficiency, Entropy and an Asymptotic Likelihood Principle," *Biometrika*, 70, pp.175-182.

Levy, Martin S., & Perng S.K. (1986). "An Optimal Prediction Function for the Normal Linear Model," *Journal of the American Statistical Association*, 81, p.196-198.

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

Liski, E.P., & Nummi, T. (1996). "Prediction in Repeated-Measures Models with Engineering Applications," *Technometrics*, 38, 25-36.

Littell, R., Milliken, G., Stroup, W., & Wolfinger, R. (1996). *SAS System for Mixed Models.* Cary, NC: SAS Institute Incorporated.

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.

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, p.139-148.

Rao, C.R. (1965). *Linear Statistical Inference and its Applications, 2nd Edition*, New York: Wiley.

Rao, C.R. (1973). *Linear Statistical Inference and its Applications, 2nd Edition*, Wiley, New York.

Rao, C.R. (1987). "Prediction of Future Observations in Growth Curve Models," *Statistical Science*, 2, 434-471.

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.

Seltzer, M. (1993). "Sensitivity Analysis for Fixed Effects in the Hierarchical Model: A Gibbs Sampling Approach," *Journal of Educational and Behavioral Statistics*, 18(3), pp.207-235.

Swamy, P.A.V.B. (1971). *Statistical Inference in Random Coefficient Regression Models.* New York: Springer Verlag.