*Communications in Statistics—Simulation and Computation*®, 34: 909–928, 2005

ISSN: 0361-0918 print/1532-4141 online
DOI: 10.1080/03610910500308248

Taylor & Francis
Taylor & Francis Group

# Linear Models

# Decomposition of Prediction Error in Multilevel Models

DAVID AFSHARTOUS[1] AND JAN DE LEEUW[2]

[1]School of Business Administration, University of Miami, Coral Gables, Florida, USA
[2]Department of Statistics, University of California, Los Angeles, California, USA

*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. Additionally, analytical results present the increase in predicted mean square error (PMSE) with respect to prediction error bias. The 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.*

**Keywords** Monte Carlo; Missing data; Multilevel model; Prediction error components.

**Mathematics Subject Classification** 62J12; 46N30.

## 1. Introduction

Consider the problem of predicting a future observable $y_{*j}$ in the $j$th 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 Ordinary Least Squares (OLS), prior, or multilevel approaches. In an earlier study, the performance of these prediction rules was assessed via a large scale simulation study (Afshartous and de Leeuw, 2005). The prediction rule employing a shrinkage estimator proved

Received November 6, 2002; Accepted June 9, 2005
Address correspondence to David Afshartous, School of Business Administration, University of Miami, Coral Gables, FL 33124-8237, USA; E-mail: afshar@miami.edu

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 and Goldstein (1998) examined the handling of educational data with students belonging to multiple groups and also the case where group membership itself is unknown.[1] Although slightly similar to Hill and Goldstein's (1998) problem, our problem centers around unknown information with respect to parameters, individual, and group covariates, and its effect on prediction.

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

### 1.1. *The Multilevel Model*

Given a hierarchical data structure, multilevel modeling represents a statistical technique that may be employed to facilitate inferences from the data. A given data point $y_{ij}$ is the $i$th case in the $j$th unit, e.g., the $i$th student in the $j$th 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 $j$th group. For a full review of the multilevel model, see Raudenbush and 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 $j$th school contains $n_j$ students. The basic multilevel model has the following level-1 model equation:

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

Following the notation in Afshartous and de Leeuw (2005), each $X_j$ has dimensions $n_j \times p$, and $r_j \sim N(0, \sigma^2\Psi_j)$, with $\Psi_j$ usually taken as $I_{n_j}$. Some or all of the level-1 coefficients, $\beta_j$, 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 $j$th group, and $u_j \sim N(0, \tau)$. As $\tau$ is not necessarily diagonal, the elements of the random

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

vector $\beta_j$ are not independent; there might exist a covariance between the slope and intercept for each regression equation.

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

$$Y_j = X_j W_j \gamma + X_j u_j + r_j \tag{3}$$

which may be viewed as a special case of the mixed linear model, with fixed effects $\gamma$ and random effects $u_j$.[2] Thus, marginally, $y_j$ has expected value $X_j W_j \gamma$ 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 $\tau$ (de Leeuw and Kreft, 1995). Thus, the full log-likelihood for the $j$th unit is

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

where $d_j = Y_j - X_j W_j \gamma$.

### 1.2. *Multilevel Prediction*

Formally, let $y_{*j}$ be the unknown outcome measure for an unsampled observation in the $j$th 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 and de Leeuw (2005) we assessed the relative performance of three prediction rules in predicting a future observable $y_{*j}$. The multilevel prediction rule is defined as follows:

$$\hat{y}_{*j} = X_{*j}\hat{\beta}_j^* \tag{5}$$

where $\hat{\beta}_j^* = W_j\hat{\gamma} + \hat{u}_j$.[3] Since $u_j$ may be interpreted in the mixed model sense as the random effect of the $j$th group, we have the commonly employed estimator of random effects from the mixed models literature (Robinson, 1991):

$$\hat{u}_j = C_j^{-1} X_j'(y_j - X_j W_j \gamma) \tag{6}$$

where $C_j = X_j' X_j + \sigma^2\tau$. 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:

$$\hat{y}_{*j} = X_{*j} W_j \hat{\gamma} + \widehat{V}_{*j}\widehat{V}_j^{-1}(y_j - X_j W_j \hat{\gamma}) \tag{7}$$

[2]See Robinson (1991) for an excellent review of estimation of fixed and random effects in the general mixed model.

[3]The estimates of $\beta_j$ may also be expressed as a linear combination of the OLS estimate $\hat{\beta}_j = (X_j' X_j)^{-1} X_j y_j$ and—given an estimate of $\gamma$—the prior estimate $W_j\hat{\gamma}$ of $\beta_j$, the weights being proportional to the estimation variance in the OLS estimate and the prior variance of the distribution of $\beta_j$.

where

$$\hat{\gamma} = \left( \sum_{j=1}^{J} W_j' X_j' \widehat{V}_j^{-1} X_j W_j \right)^{-1} \sum_{j=1}^{J} W_j' X_j' \widehat{V}_j^{-1} y_j$$

$$\widehat{V}_{*j} = \widehat{\mathrm{Cov}}(y_{*j}, y_j) = X_{*j} \hat{\tau} X_j'$$

$$\widehat{V}_j = \widehat{\mathrm{Var}}(y_j) = X_j \hat{\tau} X_j' + \hat{\sigma}^2 I$$

This representation illustrates the multilevel prediction rule as the conditional expectation of $y_{*j}$ given the data $y$.[4]

Regarding variance component estimation ($\tau$ and $\sigma^2$), maximum likelihood estimates are computed iteratively via procedures such as Fisher Scoring (Longford, 1987), iteratively reweighted generalized least squares (Goldstein, 1986), or the EM algorithm (Dempster et al., 1977). 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 and Bryk, 2002). 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, and thus the MLEs will still be negatively biased (Raudenbush and Bryk, 2002).[5]

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 article we 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.

[4]Furthermore, for the case of known $\gamma$ and known variance components, Rao (1973) showed that $\hat{y}_{*j}$ has minimum mean square error (MSE) among all linear predictors. When $\gamma$ is estimated as in Eq. 7 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).

[5]Browne and Draper (2000) have employed MCMC methods such as Gibbs sampling and Metropolis-Hasting algorithms to obtain reduced bias in variance component estimation; in order to offset the drawback of slower computational speed of the MCMC methods, they propose a hybrid approach based on likelihood and MCMC methods. 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.

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 predictor 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.[6] For higher Information States, unless otherwise specified, parameter estimates are the same as in the previous lower Information State.

Info State 2: First and Second Moments Known

$$\hat{y}_{*j} = X_{*j}W_j\gamma + V_{*j}V_j^{-1}(y_j - X_jW_j\gamma) \tag{8}$$

where

$$V_{*j} = \mathrm{Cov}(y_{*j}, y_j) = X_{*j}\tau X_j'$$
$$V_j = \mathrm{Var}(y_j) = X_j\tau X_j' + \sigma^2 I.$$

This corresponds to the ideal case where all the necessary parameters are known.[7] Thus, the parameters that are required by the multilevel prediction rule are known and estimation is 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

$$\hat{y}_{*j} = X_{*j}W_j\hat{\gamma} + V_{*j}V_j^{-1}(y_j - X_jW_j\hat{\gamma}) \tag{9}$$

where

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

Here, the coefficient $\gamma$ 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 $\gamma$ is estimated.

[6]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.

[7]Note that this formula does not produce that same predictions as simply plugging in known $X_{*j}$ and $\beta_j$ values, i.e., $\hat{y}_{*j} = X_{*j}\beta_j$. As noted earlier, in the normal case one can view this as a conditional expectation.

Info State 4: First and Second Moments Unknown

This corresponds to the situation encountered in practice, i.e., all of the model parameters must be estimated from the observed data. The relevant equations are the same as those provided in Eq. (7). 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

$$\hat{y}_{*j} = X_{*j}\overline{W}\hat{\gamma} + \widehat{V}_{*j}\widehat{V}_j^{-1}(y_j - X_j\overline{W}\hat{\gamma}) \tag{10}$$

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

$$\hat{y}_{*j} = \overline{X}_j W_j\hat{\gamma} + \widetilde{V}_{*j}\widehat{V}_j^{-1}(y_j - X_j W_j\hat{\gamma}) \tag{11}$$

where

$$\overline{X}_j = \left(1, \frac{1}{n_j}\sum_{i=1}^{n_j} X_{ij}\right)$$

$$\widetilde{V}_{*j} = \overline{X}_j\hat{\tau}X_j'.$$

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 $\gamma$. 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 (1985). 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 mis-specified. In the next sub-section, we derive analytical expressions for the predicted mean square error (PMSE) for Info States 5 and 6 in order to assess prediction error inflation over the non-missing data case.

### 2.1. *Analytical Results*

The goal is to compare the PMSEs under Info States 5 and 6 with that of Info State 2. In large sample sizes, Info State 2 is equivalent to Info State 4; thus, focus on the aforementioned differences in PMSE. Assuming normality and Info State 2 (i.e., when all fixed parameters are known), the best predictor is the regression function given previously:

$$E(y_{*j}|Y_j) = X_{*j}W_j\gamma + V_{*j}V_j^{-1}(y_j - X_jW_j\gamma) = \hat{y}. \tag{12}$$

The PMSE of the best predictor is

$$V(y_{*j}|Y_j) = \sigma_1^2 - V_{*j}V_j^{-1}V_{*j},$$

where $\sigma_1^2 = \sigma^2 + X_{*j}\tau X_{*j}'$ is the variance of the marginal distribution of $y_{*j}$. Let $\tilde{y}$ denote the expression in Eq. 12 when missing $W_j$ is replaced by $\overline{W}$, i.e., $\tilde{y}$ is the predictor when $W_j$ is missing and imputed by $\overline{W}$. Then,

$$y_{*j} - \tilde{y} = (y_{*j} - \hat{y}) + (\hat{y} - \tilde{y}) = (y_{*j} - \hat{y}) + [V_{*j}V_j^{-1}X_j - X_{*j}](\overline{W} - W_j)\gamma.$$

The second term represents the bias of $\tilde{y}$, and $\text{PMSE}(\tilde{y}) = \text{PMSE}(\hat{y}) + (\text{bias})^2$. Thus, the effect of missing $W_j$ is the bias and subsequently its contribution to PMSE.

The effect of missing $X_{*j}$ may be investigated in a similar manner. Specifically, let $\tilde{\tilde{y}}$ denote the expression in Eq. 12 when $X_{*j}$ is replaced by $\overline{X}_j$, i.e., $\tilde{\tilde{y}}$ is the predictor when $X_{*j}$ is missing and imputed by $\overline{X}_j$. Note that $V_{*j}$ will be affected as well since this term involves $X_{*j}$. As before, define $V_{*j}$ under this scenario as $\widetilde{V}_{*j} = \overline{X}_j\hat{\tau}X_j'$. Then,

$$\begin{aligned}
y_{*j} - \tilde{\tilde{y}} &= (y_{*j} - \hat{y}) + (\hat{y} - \tilde{\tilde{y}}) = (X_{*j} - \overline{X}_j)W_j\gamma + V_{*j}V_j^{-1}y_j \\
&\quad - V_{*j}V_j^{-1}X_jW_j\gamma - \widetilde{V}_{*j}V_j^{-1}y_j + \widetilde{V}_{*j}V_j^{-1}X_jW_j\gamma \\
&= (X_{*j} - \overline{X}_j)W_j\gamma + (X_{*j}\tau - \overline{X}_j\tau)X_j'V_j^{-1}y_j \\
&\quad - (X_{*j}\tau - \overline{X}_j\tau)X_j'V_j^{-1}X_jW_j\gamma \\
&= (X_{*j} - \overline{X}_j)[W_j\gamma - \tau X_j'V_j^{-1}(y_j - X_jW_j\gamma)].
\end{aligned}$$

The second term is the bias of $\tilde{\tilde{y}}$ and now involves $y_j$, and $\text{PMSE}(\tilde{\tilde{y}}) = \text{PMSE}(\hat{y}) + (\text{bias})^2$. Thus, the resulting bias for both missing level-1 ($\text{bias}_x$) and level-2 ($\text{bias}_w$) information translates into increased PMSE. Viewed side by side we have the following expressions for the bias for missing level-1 data ($\text{bias}_x$) and missing level-2 data ($\text{bias}_w$),

$$\begin{aligned}
\text{bias}_w &= [V_{*j}V_j^{-1}X_j - X_{*j}](\overline{W} - W_j)\gamma = f(W_j, \overline{W}, X_{*j}, X_j, \tau, \sigma^2, \gamma) \\
\text{bias}_x &= (X_{*j} - \overline{X}_j)[W_j\gamma - \tau X_j'V_j^{-1}(y_j - X_jW_j\gamma)] = g(y_j, X_{*j}, \overline{X}_j, W_j, \tau, \sigma^2, \gamma).
\end{aligned}$$

It is difficult to assess/visualize the relative magnitude of the bias for these two missing data scenarios. For both cases, the bias (and hence increase in PMSE) is dependent upon the accuracy of the relevant data imputation method, the population distribution of the level-1 and level-2 data, and the values of the fixed population parameters. Although the subsequent simulations provide added insight into the increase in PMSE, the results are specific to the chosen simulation design conditions. However, we may investigate whether these analytical results are in line with the simulations, i.e., whether one missing data scenario is consistently more costly across the chosen designs. The simulation design used to assess the increase in PMSE due to missing information is described in the following section.

## 3. Simulation Study Design

The design of the simulation study continues that of our previous design (Afshartous and de Leeuw, 2005), similar to that of Busing (1993) where the distribution of level-2 variance component estimates was examined. 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 2–4) and one for data (Info States 5–6). A simple 2-level model with one explanatory variable at each level and equal numbers of units per group is considered. A two-stage simulation scheme is employed. For stage one, the level-1 random coefficients are generated according to the following equations:

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

The $\gamma$'s are the fixed effects and are set equal to one. $W_j$ 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 $2 \times 2$ covariance matrix $\tau$. The two diagonal elements of $\tau$, $\tau_{00}$, and $\tau_{11}$, are equal in each design condition. The off-diagonal covariance term $\tau_{01}$ will then determine the correlation between the intercept and slope:

$$r_{u_{0j}, u_{1j}} = \frac{\tau_{01}}{(\tau_{00}\tau_{11})^{1/2}}. \tag{13}$$

The simulation design also varies the intraclass correlation $\rho$. The intraclass correlation is defined as follows:

$$\rho = \frac{\tau_{00}}{\tau_{00} + \sigma^2} \tag{14}$$

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.

Stage two of the simulation concerns the first level of the multilevel model, where observations are generated according to the following equation:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + \epsilon_{ij} \tag{15}$$

The level-1 explanatory variable, $X_{ij}$, is simulated as a standard normal random variable,[8] while the level-1 error $\epsilon_{ij}$ is a normal random variable with mean 0 and variance $\sigma^2$ specified as 0.25. As only balanced data 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.

Various sample size combinations for the number of groups ($J$) and the number of observations per group ($n$) are examined. 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 20 possible $J \times n$ combinations and twelve possible parameter specifications, yielding a total of 240 design conditions. 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.

The 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 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 PMSE 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.

The computer code for the simulations was written in XLISP-STAT[9] and the multilevel modeling was done with several altered versions of Terrace-Two.[10] 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 extend its capability.

[8]Once again, we assume that the $X$ distribution is similar across schools; this could be generalized to allow for different distributions across schools.

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

[10]An XLISP-STAT program written by James Hilden-Minton which incorporates both the EM algorithm and Fisher scoring.

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

| | $n = 5$ | | |
|---|---|---|---|
| $J$ | 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 |

## 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 various design conditions.[11] 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 11 presents the results for the $n = 5$ for various levels of $J$ across the twelve parametric design conditions.

Table 11 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. These 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 $n = 5$. Figure 1 illustrates these results via side-by-side boxplots. Note that the greatest relative penalties for the multilevel prediction rule occur in the first three boxplots, where $J = 10$. In addition, this 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).

We also note that there exists variation across the parameter design conditions (Appendix B). As each design condition is a combination of multiple parameter values, it is difficult to isolate the causes of this variation. Nevertheless, an inspection of the standard errors indicates that these differences are unlikely to be an artifact of sampling variability. Predictive performance under Design 8 seems relatively poor, whereas predictive performance under Design 9 seems relatively good. Note that Design 9 contains the lowest intraclass correlation coefficient and the highest intercept-slope correlation, while Design 8 contains the lowest intraclass correlation coefficient and a medium value for intercept-slope correlation.

[11]Only Designs #1–6 are reported due to space considerations.

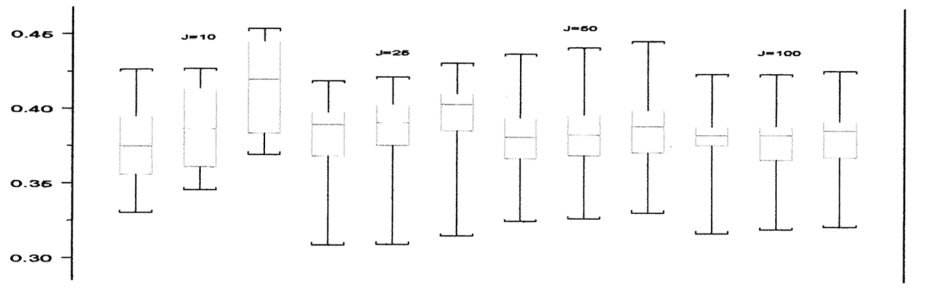


**Figure 1.** $n = 5$; $J = 10, 25, 50, 100$, MSE for Info States 2, 3, 4.

### 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 various 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 12 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.

Table 2 also 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 Figure 2.[12] 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.

The previous analytical results are in line with these simulation results. Specifically, for separate design conditions we calculated PMSE using the PMSE equations from Sec. 2.1 and compared these results with the simulations.[13] As in the simulations, there is a clear increase in the PMSE of the multilevel prediction rule for Info State 5 and 6, and PMSE for Info State 6 (missing level-1 data) is higher than that for Info State 5 (missing level-2 data). The analytical

[12]Although we only present the figure for $J = 10$ here, the corresponding figures for other values of $J$ exhibit similar patterns.

[13]Note that the level-1 and level-2 data is also generated for the analytical expressions. Similar to the simulations, we performed 100 iterations for given design conditions and calculated the analytical expressions each iteration; we compared the average of these iterations to the corresponding values for the simulations. This method was chosen such that the analytical expressions were not overly dependent on a single generation of level-1 and level-2 data values.

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

| | $n = 5$ | $n = 10$ | $n = 25$ | $n = 50$ |
|---|---|---|---|---|
| $J$ | States 4, 5, 6 | States 4, 5, 6 | States 4, 5, 6 | States 4, 5, 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 |

PMSEs are lower than the values in the simulations, reflecting the additional estimation error in the simulations. For example, in Design #1, the analytical PMSE is 19.8% lower (averaged over $J \times n$) than the simulation PMSE. When separated according to missing level-1 versus missing level-2 data, the corresponding values are 8.8 and 30.7% lower, respectively. That is, the simulations provide PMSEs closer to the analytical results for the missing level-1 data case, i.e., for the case where the PMSE is higher. As expected, the difference between the analytical results and the simulations is decreasing in both $J$ and $n$.

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. 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 Figs. 3–4. 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

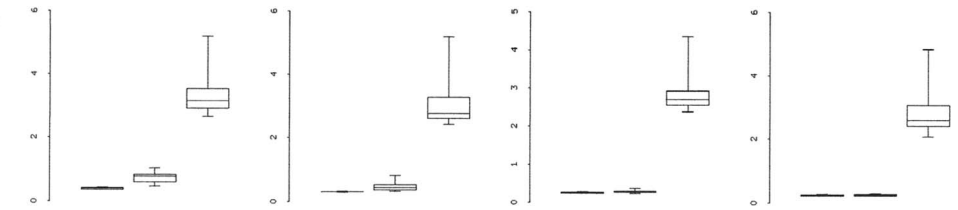

**Figure 2.** $J = 10$; $n = 5, 10, 25, 50$; MSE for Info States 4, 5, 6.

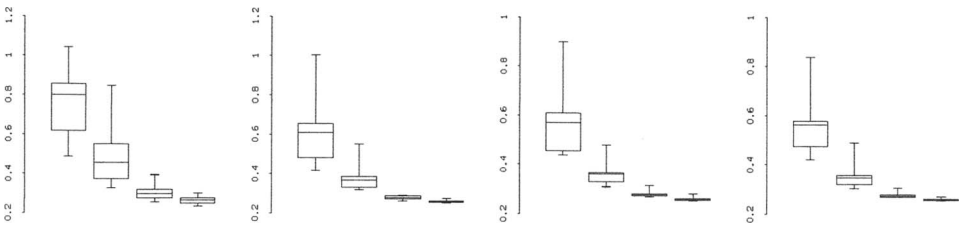

**Figure 3.** $J = 10, 25, 50, 100$; MSE for Info State 5 as $n = 5, 10, 25, 50$.

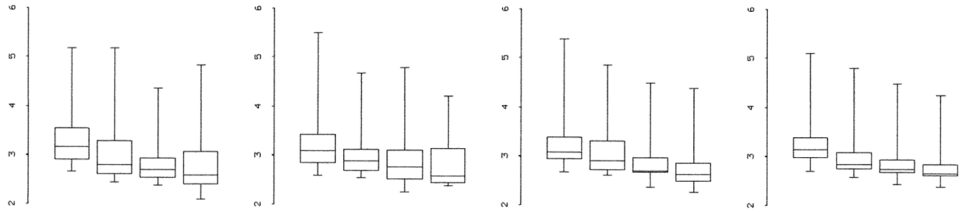

**Figure 4.** $J = 10, 25, 50, 100$; MSE for Info State 6 as $n = 5, 10, 25, 50$.

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 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 Figs. 5–6. 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.

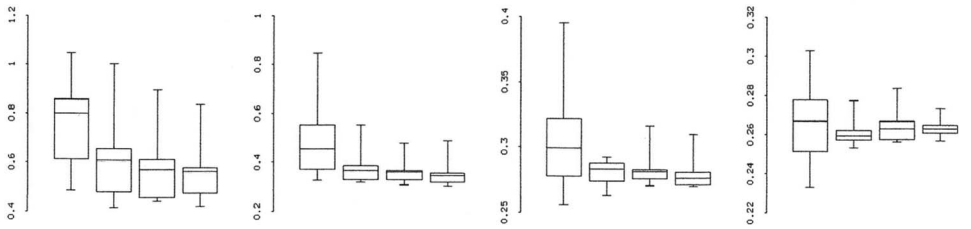

**Figure 5.** $n = 5, 10, 25, 50$; MSE for Info State 5 as $J = 10, 25, 50, 100$.

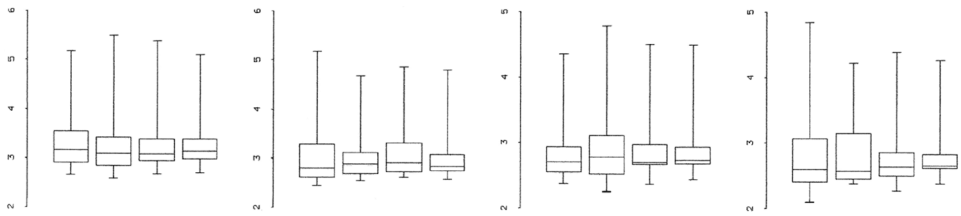

**Figure 6.** $n = 5, 10, 25, 50$; MSE for Info State 6 as $J = 10, 25, 50, 100$.

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 Fig. 7, 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-1 case (Info State 6) 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.

To be sure, the average (over all level-2 units) PMSEs, under Info State 6, are not of particular interest in assessing the cost of missing $X$ for a specific level-2 unit. This cost would be of significant practical interest to the principal of the particular school who would be interested in knowing the effect of missing $X$ for his/her school, not the average over all schools. Indeed, the distribution of $X$ for his/her school may be quite different from that of other schools. Although our simulations generated level-1 data from a standard normal distribution for all schools, in practice we are likely to see different level-1 data distributions for

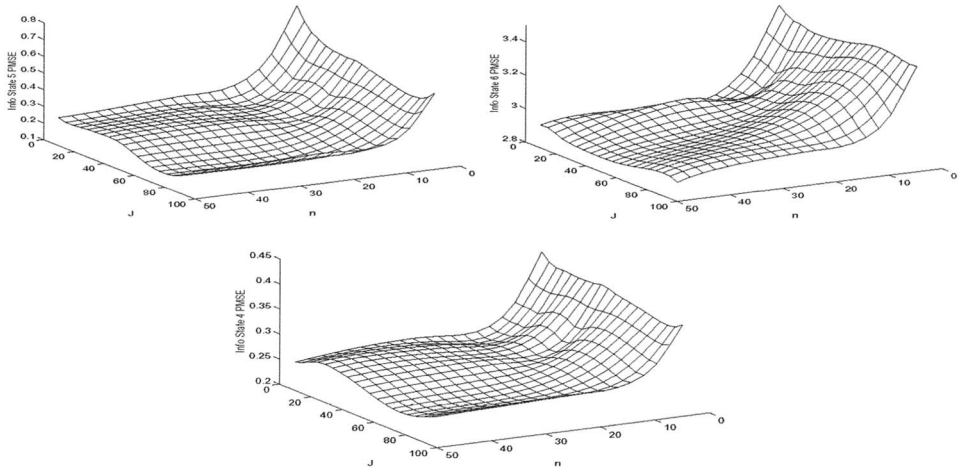


**Figure 7.** PMSE for Info States 4, 5, 6.

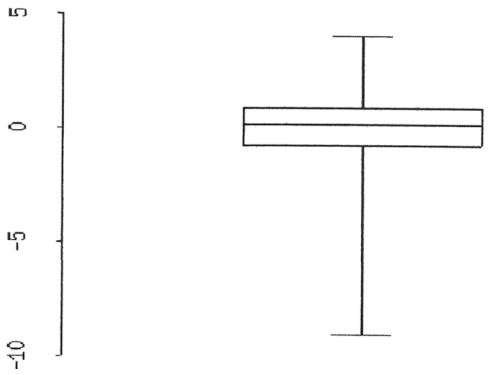


**Figure 8.** Info State 6, PE Distribution for randomly chosen group.

different schools. We examined the distribution of prediction errors for each school across the simulations.[14] Due to space limitation, in Fig. 8 we only present a randomly selected school for $J = 25, n = 10$ under design number 2. We observe a slight tendency to under-predict (median $= 0.21$), albeit we have also have some large over-predictions as exhibited by the long left tail (mean $= -0.13$). Regarding stability, the prediction errors for this group have a standard deviation of 1.9, and thus an approximate PMSE of $(1.9)^2 = 3.61$ (exact value $= 3.56$). As the average PMSE across schools equals 2.904 (Table 11), the PMSE for this school is slightly higher but still within one standard deviation (0.66). These results could be used to assess the confidence of individual predictions for this school as well.

## 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 $j$th group of a hierarchical dataset. The analytical results help assess the value of level-1 and level-2 information for prediction. Moreover, the expressions for PMSE under Info States 5 and 6 may be examined to explain the effects of design parameters on prediction error inflation. Although our simulation results indicate that missing level-1 information is more costly than missing level-2 information for prediction, such empirical results are likely to be highly dependent on the chosen design variables.

The simulation study spans the various combinations of level-1 and level-2 sample sizes and different intraclass correlation values. In addition to the components of prediction error for missing data, we also assess the cost of parameter estimation versus data imputation. For our design values, the cost of parameter estimation is very small compared to data imputation. 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

[14]Recall that each school has only one student to be predicted at each simulation iteration.

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 appears higher than that of missing level-2 information for all levels of $J$ and $n$. However, as noted above, the analytical expressions indicate that these results are highly dependent on design parameters.
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.

## Appendices

### A. Simulation Design

**Table 3**
Design numbers

| Design number | $\tau_{00}, \tau_{11}$ | $\tau_{01}$ | $r_{u_{0_j}, u_{1_j}}$ | $\rho$ |
|---|---|---|---|---|
| 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 |

## B. PE Decomposition: Info States 2, 3, 4

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

| | $n = 5$ | $n = 10$ | $n = 25$ | $n = 50$ |
|---|---|---|---|---|
| $J$ | States 2, 3, 4 | States 2, 3, 4 | States 2, 3, 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 5**
Design #2: Mean MSE for 2, 3, 4 Info States

| | $n = 5$ | $n = 10$ | $n = 25$ | $n = 50$ |
|---|---|---|---|---|
| $J$ | States 2, 3, 4 | States 2, 3, 4 | States 2, 3, 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 6**
Design #3: Mean MSE for 2, 3, 4 Info States

| | $n = 5$ | $n = 10$ | $n = 25$ | $n = 50$ |
|---|---|---|---|---|
| $J$ | States 2, 3, 4 | States 2, 3, 4 | States 2, 3, 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 7**
Design #4: Mean MSE for 2, 3, 4 Info States

| | $n = 5$ | $n = 10$ | $n = 25$ | $n = 50$ |
|---|---|---|---|---|
| $J$ | States 2, 3, 4 | States 2, 3, 4 | States 2, 3, 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 8**
Design #5: Mean MSE for 2, 3, 4 Info States

| | $n = 5$ | $n = 10$ | $n = 25$ | $n = 50$ |
|---|---|---|---|---|
| $J$ | States 2, 3, 4 | States 2, 3, 4 | States 2, 3, 4 | State 2, State 3, State 4 |
| 10 | 0.3929, 0.4103, 0.4427 | 0.3002, 0.3027, 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 9**
Design #6: Mean MSE for 2, 3, 4 Info States

| | $n = 5$ | $n = 10$ | $n = 25$ | $n = 50$ |
|---|---|---|---|---|
| $J$ | States 2, 3, 4 | States 2, 3, 4 | States 2, 3, 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 |

## C. PE Decomposition: Info States 5 and 6

**Table 10**
Design #1: Mean MSE for 5, 6 Info States

| | $n = 5$ | $n = 10$ | $n = 25$ | $n = 50$ |
|---|---|---|---|---|
| $J$ | States 5, 6 | States 5, 6 | States 5, 6 | States 5, 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 11**
Design #2: Mean MSE for 5, 6 Info States

| | $n = 5$ | $n = 10$ | $n = 25$ | $n = 50$ |
|---|---|---|---|---|
| $J$ | States 5, 6 | States 5, 6 | States 5, 6 | States 5, 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 12**
Design #3: Mean MSE for 5, 6 Info States

| | $n = 5$ | $n = 10$ | $n = 25$ | $n = 50$ |
|---|---|---|---|---|
| $J$ | States 5, 6 | States 5, 6 | States 5, 6 | States 5, 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 13**
Design #4: Mean MSE for 5, 6 Info States

| | $n = 5$ | $n = 10$ | $n = 25$ | $n = 50$ |
|---|---|---|---|---|
| $J$ | States 5, 6 | States 5, 6 | States 5, 6 | States 5, 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 14**
Design #5: Mean MSE for 5, 6 Info States

| | $n = 5$ | $n = 10$ | $n = 25$ | $n = 50$ |
|---|---|---|---|---|
| $J$ | States 5, 6 | States 5, 6 | States 5, 6 | States 5, 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 15**
Design #6: Mean MSE for 5, 6 Info States

| | $n = 5$ | $n = 10$ | $n = 25$ | $n = 50$ |
|---|---|---|---|---|
| $J$ | States 5, 6 | States 5, 6 | States 5, 6 | States 5, 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 |

## Acknowledgment

This research was supported by a grant from the National Institute for Statistical Sciences. This article was greatly enhanced by the comments of the associate editor and anonymous referee, as well as from useful discussions with Nick Afshartous and Dan Yadgar.

## References

Afshartous, D. (1995). Determination of sample size for multilevel model design. Perspectives on Statistics for Educational Research: Proceedings of the National Institute for Statistical Sciences (NISS), Technical Report #35, edited by V. S. Williams, L. V. Jones, and Ingram Oklin.

Afshartous, D. (1997). Prediction in Multilevel Models. Unpublished Ph.D., dissertation, UCLA.

Afshartous, D., de Leeuw, J. (2005). Prediction in multilevel models. *J. Educat. Behav. Statist.* 30(2):109–139.

Browne, W. J., Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics* 15:391–420.

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.

de Leeuw, J., Kreft, I. (1995). Questioning multilevel models. *J. Educat. Behav. Statist.* 20:171–189.

de Leeuw, J., Kreft, I. (2002). *Handbook of Multilevel Quantitative Analysis*. Boston, Dordrecht, London: Kluwer Academic Publishers. In Press.

de Leeuw, J. (ed.) (2005). *Handbook of Multilevel Analysis*. Springer, in preparation.

Dempster, A. P., Laird, N. M., Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* 39:1–8.

Goldberger, A. S. (1962). Best linear unbiased prediction in the general linear model. *J. Amer. Statist. Assoc.* 57:369–375.

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

Harville, D. A. (1976). Extension of the Gauss Markov theorem to include the estimation of random effects. *Ann. Statist.* 4:384–396.

Harville, D. A. (1985). Decomposition of prediction error. *J. Amer. Statist. Assoc.* 80:132–138.

Hill, P. W., Goldstein, H. (1998). Multilevel modelling of educational data with cross-classification and missing identification for units. *J. Educat. Behav. Statist.* 23:117–128.

Longford, N. T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika* 74:817–827.

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

Raudenbush, S., Bryk, A. (2002). *Heirarchical Linear Models*. Newbury Park: Sage Publications.

Robinson, G. K. (1991). That BLUP is a good thing. *Statist. Sci.* 6:15–51.