# Optional ERIC Coversheet — Only for Use with U.S. Department of Education Grantee Submissions

This coversheet should be completed by grantees and added to the PDF of your submission if the information required in this form **is not included on the PDF to be submitted**.

**INSTRUCTIONS**

- Before beginning submission process, download this PDF coversheet if you will need to provide information not on the PDF.
- Fill in all fields—information in this form **must match** the information on the submitted PDF and add missing information.
- Attach completed coversheet to the PDF you will upload to ERIC [use Adobe Acrobat or other program to combine PDF files]—do not upload the coversheet as a separate document.
- Begin completing submission form at https://eric.ed.gov/submit/ and upload the full-text PDF with attached coversheet when indicated. Your full-text PDF will display in ERIC after the 12-month embargo period.

## GRANTEE SUBMISSION REQUIRED FIELDS

**Title of article, paper, or other content**

Maximum Likelihood Estimation of Hierarchical Linear Models from Incomplete Data: Random Coefficients, Statis

**All author name(s) and affiliations on PDF. If more than 6 names, ERIC will complete the list from the submitted PDF.**

| Last Name, First Name | Academic/Organizational Affiliation | ORCID ID |
|---|---|---|
| Shin, Yongyun | Virginia Commonwealth University | yshin@vcu.edu |
| Raudenbush, Stephen | University of Chicago | |
| | | |
| | | |
| | | |
| | | |

**Publication/Completion Date—**(if *In Press,* enter year accepted or completed) June 24, 2023

**Check type of content being submitted and complete one of the following in the box below:**

- ☑ If article: Name of journal, volume, and issue number if available
- ☐ If paper: Name of conference, date of conference, and place of conference
- ☐ If book chapter: Title of book, page range, publisher name and location
- ☐ If book: Publisher name and location
- ☐ If dissertation: Name of institution, type of degree, and department granting degree

Journal of Computational and Graphical Statistics

**DOI or URL to published work** (if available) https://doi.org/10.1080/10618600.2023.2234414

**Acknowledgement of Funding**— Grantees should check with their grant officer for the preferred wording to acknowledge funding. If the grant officer does not have a preference, grantees can use this suggested wording (adjust wording if multiple grants are to be acknowledged). Fill in Department of Education funding office, grant number, and name of grant recipient institution or organization.

"This work was supported by U.S. Department of Education **[Office name]** the Institute of Education S through **[Grant number]** R305D2100 to **Institution]** Virginia Comm. The opinions expressed are those of the authors and do not represent views of the **[Office name]** the Institute or the U.S. Department of Education.

# Maximum Likelihood Estimation of Hierarchical Linear Models from Incomplete Data: Random Coefficients, Statistical Interactions, and Measurement Error

Yongyun Shin

Virginia Commonwealth University

830 East Main Street

Richmond, VA 23298-0032

yshin@vcu.edu

and

Stephen W. Raudenbush

University of Chicago

sraudenb@uchicago.edu

**Abstract**

We consider two-level models where a continuous response $R$ and continuous covariates $C$ are assumed missing at random. Inferences based on maximum likelihood or Bayes are routinely made by estimating their joint normal distribution from observed data $R_{obs}$ and $C_{obs}$. However, if the model for $R$ given $C$ includes random coefficients, interactions, or polynomial terms, their joint distribution will be nonstandard. We propose a family of unique factorizations involving selected "provisionally known random effects" $u$ such that $h(R_{obs}, C_{obs}|u)$ is normally distributed and $u$ is a low-dimensional normal random vector; we approximate $h(R_{obs}, C_{obs}) = \int h(R_{obs}, C_{obs}|u)g(u)du$ via adaptive Gauss-Hermite quadrature. For polynomial models, the approximation is exact but, in any case, can be made as accurate as required given sufficient computation time. The model incorporates random effects as explanatory variables, reducing bias due to measurement error. By construction, our factorizations solve problems of compatibility among fully conditional distributions that have arisen in Bayesian imputation based on the Gibbs Sampler. We spell out general rules for selecting $u$, and show that our factorizations can support fully compatible Bayesian methods of imputation using the Gibbs Sampler.

KEY WORDS: Provisionally known random effects; the EM algorithm; adaptive Gauss Hermite quadrature; compatibility; missing at random.

# 1 Introduction

Large-scale surveys and experiments within the social and health sciences frequently meet four conditions that supply the focus of this article. First, the data typically have a hierarchical structure, with respondents nested within local organizational units such as schools and hospitals or repeated measures nested within persons. Second, missing data are pervasive. Third, partially observed covariates may be measured with error. Finally, the covariates of interest may have random coefficients, statistical interactions or polynomial terms.

These characteristics have received some attention in recent methodological research. A popular approach conceives the response variables and partially observed covariates as outcomes within a multivariate, hierarchical linear model (HLM) under the assumption that the data are missing at random (MAR; Rubin, 1976), an assumption often thought reasonable given a presumably rich set of covariates (Schafer and Yucel, 2002; Goldstein et al., 2014). Missing values are imputed from their posterior predictive distribution of the missing values using what has been termed a "fully conditional specification" (FCS) using the Gibbs sampler. FCS requires the analyst to impute missing values for each variable subject to missingness, conditional on all other unknowns. This approach has been shown to function well when the process generating the joint distribution of unknowns is reasonably assumed multivariate normal. Under the four conditions just described, however, multivariate normality is not possible even if the separate conditional distributions are normal. A concern involves the bias that can arise from the incompatibility between the multiple normal conditional distributions that generate the imputations and the assumed joint distribution of the observed data (Erler et al., 2016; Enders et al., 2016; Enders et al., 2020).

In this article, we propose to address the problem of compatibility (Arnold and Press, 1989; Liu et al., 2014; Bartlett et al., 2015) within the framework of maximum likelihood (ML) estimation under the ignorable missing data assumption that data are MAR and that the parameter spaces for the multivariate HLM and the missing data mechanism are distinct (Rubin 1976; Little and Rubin 2002).

We will first review currently popular methods of inference for normal-theory multilevel models given incomplete data and show how to modify such approaches in which random effects often become explanatory variables, allowing us to model fallible measurement as a source of incomplete data. We'll then describe our approach when incompletely observed covariates have random coefficients or polynomial terms, including statistical interactions. If a carefully selected subset of these random effects is conditionally assumed known, the model of interest can plausibly follow a normal-theory specification. The "provisionally known random effects (PKREs)" thus selected can be integrated out of the likelihood using numerical methods. Capitalizing on the invariance properties of ML estimates (MLE), we show that our re-parameterized model can be translated back to the original model's parameter space. Auxiliary variables are introduced to increase the robustness of the MAR assumption. We will illustrate application of the model by studying income inequality and mathematics achievement in US elementary schools. Although we base our case study on MLE, the likelihood factorization we propose can readily be implemented within a Bayesian approach with assurance of compatibility between conditional distributions and the joint distribution of observed data elements. We elucidate general rules for selecting low-dimensional "provisionally constant" random effects for a general class of models involving random coefficients or polynomial terms, including interactions.

Section 2 reviews estimation of normal-theory hierarchical models from data MAR. Section 3 explains how to estimate an analytic hierarchical model with random coefficients and cross-level interaction effects given data MAR via a PKRE. Section 4 describes estimation of the analytic model using auxiliary covariates to strengthen the MAR assumption. Section 5 extends the model with polynomial terms including within-level interactions and spells out rules for selecting provisionally constant random effects. Section 6 illustrates analysis of income inequality in math achievement. Section 7 evaluates our estimators by simulation. Finally, section 8 discusses the limitations and extensions.

# 2 Inference for Multivariate Normal HLMs from Incomplete Data Using Random Effects as Predictors

We begin with a review of the normal theory case. Following Schafer and Yucel (2002), our scientific interest focuses on the regression of a response variable $R^*$ on covariates $C^*$. The elements of $R^*$ and $C^*$, which are continuously measured, are partially observed. Ours is a two-level HLM (Lindley and Smith, 1972; Dempster et al., 1981) in which the response variable $R^*$ is a characteristic of "level-1 units" (e.g., students) who are clustered within level-2 units (e.g., schools). In contrast, the covariates may be characteristics of either level-1 units or level-2 units. In longitudinal studies, the level-1 units might be repeated occasions of measurements clustered within persons at level 2. Our scientist is primarily interested in the conditional distribution $f_1(R^*|C^*)$ assumed normally distributed. However, to account for the missing values in $C^*$, we propose a normal linear model $f_2(C^*)$. Because of missing data, we cannot separately estimate the parameters of $f_1$ and $f_2$ without discarding the cases with missing values on any element of $R^*$ or $C^*$. It is well known that a procedure that analyzes only the cases with complete data is prone to bias and/or loss of efficiency (Little and Rubin, 2002). The cost can be particularly high when a missing item of $C^*$ varies at level 2, in which case all of the level-1 units in the level-2 unit having missing values are discarded along with that level-2 unit itself.

To make efficient inference possible, and following Schafer and Yucel (2002), we compose each outcome vector $Y^* = [R^*\ C^{*T}]^T$ and write a multivariate HLM $h(Y^*)$

$$Y^* \quad = \quad X^*\alpha + Z^*b + r^* \sim N(X^*\alpha, V^* = Z^*\Omega Z^{*T} + \Sigma^*) \tag{1}$$

where $b \sim N(0, \Omega)$ and $r^* \sim N(0, \Sigma^*)$. Here, $X^*$ and $Z^*$ are composed of completely observed covariates; $\alpha$ is a vector of fixed regression coefficients while $b$ and $r^*$ are independent random effects that vary at levels 2 and 1, respectively. We partition the complete data (CD) $Y^*$ into

components $Y^* = (Y_{obs}, Y_{mis})$. In particular, if $Y^*$ is $N$ by 1 but we observe only $M(\leq N)$ elements $Y_{obs} = Y$ of $Y^*$, we construct an $M$-by-$N$ matrix $O$ in which every row contains a single entry equal to unity indicating which value of $Y^*$ that is observed. All other entries in the same row are 0. Our model for the observed $Y$ is therefore

$$Y \;=\; X\alpha + Zb + r \sim N(X\alpha, V = Z\Omega Z^T + \Sigma) \tag{2}$$

where $Y = OY^*$, $X = OX^*$, $Z = OZ^*$, $r = Or^*$ and $\Sigma = O\Sigma^* O^T$. Assuming data MAR, we can make efficient estimates of the parameters $\theta = (\alpha, \Omega, \Sigma^*)$ using ML or Bayes inference from the observed data according to Equation (1). The most common method in recent literature is a Bayesian approach based on multiple imputation (MI) that we will consider in the final section of this article (see "Discussion"). However, Schafer and Yucel (2002) showed how one can obtain MLE using the EM algorithm and use the estimates for MI. Shin and Raudenbush, SR hereafter, (2007) showed how to recover MLE for the analytic model $f_1(R^*|C^*)$ by constructing model (1) carefully and estimating the model without a need for imputations. This approach allows some components of $Y^*$ to vary at level 2.

## 2.1 Estimation via the EM algorithm

The EM algorithm (Dempster et al. 1977) requires evaluation at each iteration $m+1$ of the conditional expected CD score given the observed data and parameter estimates at iteration $m$. To find the CD score, our model for the CD $Y^*$ is a multivariate HLM (1). Write model (1) at the level of cluster $j$ and let $\varphi$ be an arbitrary scalar element of $(\Omega, \Sigma^*)$. The CD score equations are well known (c.f., Raudenbush and Bryk 2002, Chapter 14):

$$\begin{aligned} S_{\alpha,CD} &= \sum_j X_j^{*T} V_j^{*-1} (Y_j^* - X_j^* \alpha), \\ S_{\varphi,CD} &= \frac{1}{2} \sum_j \left( \frac{dvec(V_j^*)}{d\varphi} \right)^T (V_j^{*-1} \otimes V_j^{*-1}) vec\left[ (Y_j^* - X_j^*\alpha)(Y_j^* - X_j^*\alpha)^T - V_j^* \right]. \end{aligned} \tag{3}$$

The conditional expected score equations given the observed $Y_j$ thus clearly depend on the conditional mean and variance of $Y_j^*$ which we readily derive from the fact that $Y_j^{*(m+1)}|Y_j, \hat{\theta}^{(m)} \sim N\left(\hat{Y}_j^{*(m+1)}, \hat{V}_j^{*(m+1)}\right)$ given iteration-$m$ estimates $\hat{\theta}^{(m)} = (\hat{\alpha}^{(m)}, \hat{\Omega}^{(m)}, \hat{\Sigma}^{*(m)})$ (SR 2007) where

$$\begin{aligned} \hat{Y}_j^{*(m+1)} &= X_j^* \hat{\alpha}^{(m)} + \hat{V}_j^{*(m)} O_j^T \hat{V}_j^{(m)-1} (Y_j - X_j \hat{\alpha}^{(m)}), \\ \hat{V}_j^{*(m+1)} &= \hat{V}_j^{*(m)} - \hat{V}_j^{*(m)} O_j^T \hat{V}_j^{(m)-1} O_j \hat{V}_j^{*(m)}. \end{aligned} \tag{4}$$

## 2.2 Example 1: Contextual Effects Model

We apply this framework to the "contextual effects model" (Willms, 1986) for the study of income inequality in the mathematics achievement of US elementary school children. This model decomposes the association between family income and educational achievement into a within-school component and a between-school component. In its simplest form, the model is typically written as

$$R_{ij} = \gamma_{00} + \gamma_{10}(C_{ij} - \bar{C}_j) + \gamma_{01}(\bar{C}_j - \bar{C}) + u_{0j} + e_{ij} \tag{5}$$

where $R_{ij}$ is a measure of math achievement for student $i$ in school $j$, $C_{ij}$ is a measure of the income of that child's parents, $\bar{C}_j$ is the sample mean of $C_{ij}$ within school $j$, and $\bar{C}$ is the overall sample mean income for $i = 1, \cdots, n_j$ and $j = 1, \cdots, J$. Here $\gamma_{10}$ is known as the "within-school coefficient" while $\gamma_{01}$ is the "between school" coefficient; and $u_{0j}$ and $e_{ij}$ are independent, normally distributed random effects that vary at levels 2 and 1, respectively. Of interest is the "contextual component" $\gamma_c = \gamma_{01} - \gamma_{10}$ which, if positive, suggests that attending a school with high-income peers predicts elevated achievement net of the contribution of a student's family income.

Two problems arise in the conventional analysis (SR, 2010). First, past analyses have treated parental income as completely observed when, in fact, most surveys report substantial fractions of missing data on income. Second, the sample mean $\bar{C}_j$ will be a noisy proxy for

the actual mean income of parents in a school if the sample size per school is modest, as is the case in most US national surveys. Using random effects as explanatory variables within the MAR framework addresses both issues. Reflecting that income is partially observed, we propose the CD model

$$R^*_{ij} = \gamma_{00} + \gamma_{10}\epsilon^*_{ij} + \gamma_{01}\nu_j + u_{0j} + e^*_{ij}, \qquad C^*_{ij} = \delta + \nu_j + \epsilon^*_{ij} \tag{6}$$

where $\nu_j$ and $\epsilon^*_{ij}$ are independent, normally distributed random effects at levels 2 and 1, respectively. The joint distribution of $R^*_{ij}$ and $C^*_{ij}$ may be written

$$\begin{bmatrix} R^*_{ij} \\ C^*_{ij} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \gamma_{00} \\ \delta \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \gamma_{01}\nu_j + u_{0j} \\ \nu_j \end{bmatrix} + \begin{bmatrix} \gamma_{10}\epsilon^*_{ij} + e^*_{ij} \\ \epsilon^*_{ij} \end{bmatrix}. \tag{7}$$

Stacking the equations within level-2 unit $j$, we have the general form of the model

$$Y^*_j = X^*_j\alpha + Z^*_j b_j + r^*_j \sim N(X^*_j\alpha, V^*_j = Z^*_j \Omega Z^{*T}_j + I_{n_j} \otimes \Sigma^*) \tag{8}$$

for $X^*_j = Z^*_j = 1_{n_j} \otimes I_2$, $b_j \sim N(0, \Omega)$ and $r^*_j \sim N(0, I_{n_j} \otimes \Sigma^*)$ where we denote $1_m$ as a vector of $m$ unities and $I_m$ an $m$-by-$m$ identity matrix for a positive integer $m$. The HLM score equations for $\theta$ are familiar (Raudenbush and Bryk, 2002, Chapter 14; SR 2010) and will therefore not be elaborated here.

## 2.3 Compatibility

A set of conditional distributions $f_1(R^*|C^*, \theta_1)$ and $f_2(C^*|R^*, \theta_2)$ is said to be compatible if there exist a joint distribution $\{h(Y^*|\theta) : \theta \in \Theta\}$ and surjective maps $\{t_j : \Theta \to \Theta_j : j = 1, 2\}$ such that for each $j$, $\theta_j \in \Theta_j$ and $\theta \in t_j^{-1}(\theta_j) = \{\theta : t_j(\theta) = \theta_j\}$, we have $f_1(R^*|C^*, \theta_1) = h(R^*|C^*, \theta)$ and $f_2(C^*|R^*, \theta_2) = h(C^*|R^*, \theta)$ (Liu et al. 2014). Assuming a prior $p(\theta)$, Schafer and Yucel (2002) developed the Gibbs sampler based on multivariate normal (MN) $h(Y^*|\theta)$

that is compatible with an analytic HLM $f_1(R^*|C^*, \theta_1)$ when $C^*$ is linearly associated with $R^*$. The MN $h(Y^*|\theta)$, however, cannot be compatible with $f_1(R^*|C^*, \theta_1)$ when $C^*$ has non-linear effects (Kim et al. 2015; Enders et al. 2020). Goldstein et al (2014) and Enders et al. (2020) factored $h(Y^*|\theta) = f_1(R^*|C^*, \theta_1)f_2(C^*|\theta_2')$ and estimated a compatible HLM $f_1(R^*|C^*, \theta_1)$ with the non-linearities by the Gibbs sampler via a Metropolis algorithm. We extend the HLM further with the nonlinear effects of $C^*$ that includes random effects as latent covariates. We estimate $h(Y^*|\theta)$ efficiently by ML via a PKRE and translate the estimates to the ML estimates of the compatible HLM as explained in the next section.

# 3 Coping with Random Coefficients and Interactions

Our scientific focus is again on $f_1(R^*|C^*, \nu)$ assumed normal in distribution as in Equations (6). However, the model now includes elements of $C^*$ that have non-linearities including random coefficients, polynomial terms, or interactions. In this case, even if we can reasonably assume that $f_2(C^*)$ is normal, the joint distribution $h(Y^*)$ cannot be normal.

Recent research on the problem of non-linearities has focused primarily on two widely used methods of analysis of multilevel incomplete data. Perhaps the most popular method of imputation for HLMs under the MAR assumption imputed missing values by a series of sequential univariate regression models (Raghunathan et al. 2001), which is also known as MI by fully conditional specification or FCS (van Buuren et al. 2006). However, these conditionals will not be compatible with the joint distribution of interest in the presence of the non-linearities of interest in this article, as shown by Enders et al. (2016) and Enders et al. (2020). The main alternative approach estimates the joint distribution of the outcome and covariates subject to missingness. Missing values may be imputed based on their fully conditional distributions (Liu et al. 2000; Schafer and Yucel 2002; Goldstein et al. 2009) and by maximum likelihood (SR 2007, 2010; Ren and Shin 2016). These normal-theory models were not designed to handle the non-linearities of interest here. By means of a Gibbs

sampler via the Metropolis Hastings algorithm, Goldstein et al. (2014) imputed missing values of a response and covariates including interaction and polynomial terms having fixed effects in a multilevel model where covariates and response may be continuous or categorical. Similarly, Erler et al. (2016) took the sequential fully Bayesian approach (Ibrahim et. al. 2002) that expresses the joint distribution of variables MAR, including the outcome, into a series of univariate conditional models to handle missing values of cluster-level continuous and discrete covariates having fixed effects. Erler et al. (2019) extended the approach to imputing missing values of level-1 covariates having fixed effects. Enders et al. (2020) showed, however, that these approaches do not guarantee compatibility when the partially observed covariates have random coefficients. Lacking a formal model for the joint distribution of interest, these approaches appear to fall short of ensuring compatibility.

## 3.1 Factorization of the Likelihood Based on Provisionally Known Random Effects

To cope with nonlinearities induced by partially observed covariates having random coefficients, polynomials, or interaction terms, we model the joint distribution $h(Y^*|\nu)p(\nu)$ induced by the scientific model of interest $f_1(R^*|C^*,\nu)$ and the model for the covariates, $f_2(C^*|\nu)p(\nu)$. The problem is similar to the problem of estimation of generalized linear mixed models (Hedeker and Gibbons, 1994; Raudenbush et al., 2000). Using the notation of Equation (2), we must evaluate

$$h(Y) = \int h(Y|b)p(b)db. \tag{9}$$

The integral just defined does not have closed form in the presence of non-linearities and must be approximated numerically. To facilitate the approximation, we choose a PKRE, call

it $u$ such that

$$h(Y|u) = \int h(Y|b,u)p(b|u)db \tag{10}$$

is a normal HLM discussed in Section 2. The problem of approximation is then to evaluate

$$h(Y) = \int h(Y|u)g(u)du. \tag{11}$$

The computational challenge is to select $u$ such that the dimension of the analytic integral (10) is maximized while the dimension of numerical approximation (11) is minimized.

## 3.2 Example 2: A Partially observed covariate having a random coefficient

We return to our example of income inequality in US elementary schools. A common finding in multilevel studies of educational achievement is that relationship between student socioeconomic background and achievement varies from school to school (Raudenbush and Bryk, 1986). This variation could reflect variation in school organization, composition, and resources. To assess the variation in income inequality within schools, we expand the contextual model (6) to allow for a random coefficient

$$R^*_{ij} = (\gamma_{00} + u^*_{0j}) + (\gamma_{10} + u^*_{1j})\epsilon^*_{ij} + e^*_{ij}, \qquad C^*_{ij} = \delta + \nu_j + \epsilon^*_{ij} \tag{12}$$

where $u^*_{0j} \sim N(0, \tau_{00})$, $u^*_{1j} \sim N(0, \tau_{11})$, $\nu_j \sim N(0, \tau_{\nu\nu})$, $cov(u^*_{0j}, u^*_{1j}) = \tau_{01}$, $cov(u_{0j}, \nu_j) = \tau_{0\nu}$ and $cov(u^*_{1j}, \nu_j) = \tau_{1\nu}$. This is model (6) for $u^*_{0j} = \gamma_{01}\nu_j + u_{0j}$ if $\tau_{11} = 0$. Let $var(u^*_j) = \tau$ for $u^*_j = [u^*_{0j}\ u^*_{1j}\ \nu_j]^T$. Level-1 random effects $e^*_{ij} \sim N(0, \sigma^2)$ and $\epsilon^*_{ij} \sim N(0, \sigma_{cc})$ are assumed independent of each other and of the level-2 random effects. We denote the parameters of the CD model (12) as $\theta^*_{(12)} = (\gamma_{00}, \gamma_{10}, \tau, \sigma^2, \delta, \sigma_{cc})$.

Clearly $R^*_{ij}$ cannot be marginally normal in distribution because of the multiplication of

the two normal random effects $u^*_{1j}$ and $\epsilon^*_{ij}$. Our strategy is to select one of these two random effects to be considered "provisionally known;" we choose $u^*_{1j}$ for this purpose because it has lower dimension varying across schools than does $\epsilon^*_{ij}$ which varies across students within each school. Therefore, we write

$$\left[\begin{array}{c} u^*_{0j} \\ \nu_j \end{array}\right] \Big| u^*_{1j} \;\sim\; N\left(\left[\begin{array}{c} \alpha_{0|1} \\ \alpha_{\nu|1} \end{array}\right] u^*_{1j}, \Omega = \left[\begin{array}{cc} \tau_{00|1} & \tau_{0\nu|1} \\ \tau_{\nu 0|1} & \tau_{\nu\nu|1} \end{array}\right]\right) \tag{13}$$

where $\alpha_{k|1} = \tau_{k1}\tau_{11}^{-1}$ and $\tau_{kk'|1} = \tau_{kk'} - \alpha_{k|1}\tau_{11}\alpha_{k'|1}$ for $k, k' = 0, \nu$. We can therefore write the "provisional" joint model $h(Y^*_j|u^*_{1j})$ for $Y^*_j = [Y^{*T}_{1j} \cdots Y^{*T}_{n_j j}]^T$ as

$$Y^*_j = X^*_j\alpha + Z^*_j b_j + r^*_j \;\sim\; N(X^*_j\alpha, V^*_j = Z^*_j\Omega Z^{*T}_j + \Psi^*_j) \tag{14}$$

where $X^*_j = 1_{n_j} \otimes [I_2 \; I_2 u^*_{1j}]$, $\alpha = [\gamma_{00} \; \delta \; \alpha_{0|1} \; \alpha_{\nu|1}]^T$, $Z^*_j = 1_{n_j} \otimes I_2$, $b_j = [u^*_{0j} - \alpha_{0|1}u^*_{1j} \; \nu_j - \alpha_{\nu|1}u^*_{1j}]^T \sim N(0, \Omega)$ and $r^*_j = [r^{*T}_{1j} \cdots r^{*T}_{n_j j}]^T \sim N(0, \Psi^*_j = I_{n_j} \otimes \Sigma^*_j)$ for

$$r^*_{ij} = \left[\begin{array}{cc} 1 & \gamma_{10} + u^*_{1j} \\ 0 & 1 \end{array}\right] \left[\begin{array}{c} e^*_{ij} \\ \epsilon^*_{ij} \end{array}\right], \; \Sigma^*_j = \left[\begin{array}{cc} (\gamma_{10} + u^*_{1j})^2\sigma_{cc} + \sigma^2 & (\gamma_{10} + u^*_{1j})\sigma_{cc} \\ (\gamma_{10} + u^*_{1j})\sigma_{cc} & \sigma_{cc} \end{array}\right]. \tag{15}$$

The CD score is therefore familiar; see Equations (3). To complete the EM algorithm to estimate $h(Y^*_j|u^*_{1j})\phi(u^*_{1j}; 0, \tau_{11})$, we will maximize the likelihood $L(\theta) = \prod_{j=1}^J h(Y_j)$ for

$$h(Y_j) = \int h(Y_j|u^*_{1j})\phi(u^*_{1j}; 0, \tau_{11})du^*_{1j}, \quad h(Y_j|u^*_{1j}) \sim N(X_j\alpha, V_j = Z_j\Omega Z^T_j + \Psi_j) \tag{16}$$

where $h(Y_j|u^*_{1j})$ is from Equation (14), $X_j = O_j X^*_j$, $Z_j = O_j Z^*_j$ and $\Psi_j = O_j\Psi^*_j O^T_j = \oplus_{i=1}^{n_j}\Sigma_{ij}$ for $O_j = \oplus_{i=1}^{n_j} O_{ij}$ and $\Sigma_{ij} = O_{ij}\Sigma^*_j O^T_{ij}$. We use adaptive Gauss-Hermite Quadrature (AGHQ) to numerically approximate integral (16) (Naylor and Smith 1982; Pinheiro and Bates 1995; Rabe-Hesketh et al. 2002).

An additional AGHQ step is needed to complete the E step by evaluating the expectation

of a CD score component $S_{CDj}$ of cluster $j$ from Equations (3)

$$E(S_{CDj}|Y_j) = \int\int S_{CDj} f(Y_j^*|Y_j, u_{1j}^*) g(u_{1j}^*|Y_j) dY_{misj} du_{1j}^* \tag{17}$$

where $f(Y_j^*|Y_j, u_{1j}^*)$ has the familiar form of the empirical Bayes posterior normal density with the means and covariance matrix in Equations (4). We use AGHQ to approximate the outer integral with respect to the univariate random effect, $u_{1j}^*$; see Appendix C for detail. Because $g(u_{1j}^*|Y_j)$ is nonstandard, we use the Bayes theorem to find

$$g(u_{1j}^*|Y_j) \quad = \quad \frac{h(Y_j|u_{1j}^*)\phi(u_{1j}^*; 0, \tau_{11})}{h(Y_j)}. \tag{18}$$

By the invariance property of MLE, we translate the MLE of the "provisional" parameters $\theta = (\alpha, \Omega, \gamma_{10}, \sigma_{cc}, \sigma^2, \tau_{11})$ of model (14) to a one-to-one transformation $\hat{\theta}_{(12)}^*$ via model (13).

### 3.3 Example 3: Cross-level interaction effects involving partially-observed covariates

Following Lee and Bryk (1989), we wish to extend the contextual effects model in two ways to allow the level-1 covariate: (i) to have random coefficients as in the previous section; and (ii) to interact with the level-2 covariate. We therefore write the model

$$R_{ij}^* = (\gamma_{00} + \gamma_{01}\nu_j + u_{0j}) + (\gamma_{10} + \gamma_{11}\nu_j + u_{1j})\epsilon_{ij}^* + e_{ij}^*, \qquad C_{ij}^* = \delta + \nu_j + \epsilon_{ij}^* \tag{19}$$

where $u_{0j}$ and $u_{1j}$ are, as before, bivariate normal, but conditional on $\nu_j$ with variances $\tau_{00|\nu}$ and $\tau_{11|\nu}$, respectively, and covariance $\tau_{01|\nu}$. Other random effects are as defined in model (12). The parameters are $\theta_{(19)}^* = (\gamma_{00}, \gamma_{01}, \gamma_{10}, \gamma_{11}, \tau_{00|\nu}, \tau_{01|\nu}, \tau_{11|\nu}, \sigma^2, \delta, \tau_{\nu\nu}, \sigma_{cc})$.

This model involves two products of normal theory random effects: $\nu_j\epsilon_{ij}^*$ and $u_{1j}\epsilon_{ij}^*$. Using the logic of the last section, we wish to choose a PKRE such that, conditional on that effect, our CD joint model will be a normal-theory HLM. The question is how to choose

this random effect. We might provisionally hold both $\nu_j$ and $u_{1j}$ constant, but we would prefer to minimize the dimension of the provisionally constant random effects such that the computational burden of numerical approximation is a minimum. Alternatively, we could provisionally hold $\epsilon^*_{ij}$ constant. But while $\epsilon^*_{ij}$ is a scalar, it varies across all level-1 units, which could be a very large set. Instead, we choose to provisionally hold constant the scalar level-2 random effect

$$u^*_{1j} = \gamma_{11}\nu_j + u_{1j}. \tag{20}$$

In addition, we define $u^*_{0j} = \gamma_{01}\nu_j + u_{0j}$ to represent the model parsimoniously as $R_{ij} = (\gamma_{00} + u^*_{0j}) + (\gamma_{10} + u^*_{1j})\epsilon^*_{ij} + e_{ij}$. This model is therefore equivalent to model (12) to imply Equations (13)-(18) and a one-to-one correspondence between $\theta^*_{(12)}$ and $\theta^*_{(19)}$.

Consequently, the CD model (12) is a one-to-one transformation of the provisional model $h(Y^*_{ij}|u^*_{1j})\phi(u^*_{1j}; 0, \tau_{11})$ in Equation (14) by distribution (13) and, also, of $h(Y^*_{ij}|\nu_j)\phi(\nu_j; 0, \tau_{\nu\nu}) = f_1(R^*_{ij}|C^*_{ij}, \nu_j)f_2(C^*_{ij}|\nu_j)\phi(\nu_j; 0, \tau_{\nu\nu})$ in Equations (19) by Equation (20). We choose to estimate joint model (14) via the PKRE for efficient computation that is guaranteed to be compatible with the scientific model $f_1(R^*_{ij}|C^*_{ij}, \nu_j)$ by the one-to-one correspondence. Whereas scientific interest focuses on $\theta^*_{(19)}$, we will be estimating the parameters $\theta = (\alpha, \Omega, \gamma_{10}, \sigma_{cc}, \sigma^2, \tau_{11})$ of the provisional joint model (14). We then exploit the invariance property of MLE again, translating the MLE of $\theta$ back to those of $\theta^*_{(19)}$.

The standard approach of replacing $\nu_j$ and $\epsilon^*_{ij}$ with $\bar{C}^*_j - \bar{C}^*$ and $C^*_{ij} - \bar{C}^*_j$, respectively, in model (19) produces biased estimation of $(\gamma_{01}, \gamma_{11}, \tau_{00|\nu}, \tau_{01|\nu}, \tau_{11|\nu})$ even if $R^*_{ij}$ and $C^*_{ij}$ were fully observed; see Appendix A. Model (19) can readily incorporate multiple covariates having random effects and also having multiple cross-level interactions. Moreover, it is straightforward to include covariates having fixed coefficients. This is important given the need to add auxiliary information to strengthen the robustness of the MAR assumption.

# 4 Auxiliary Covariates

Our focus is on estimation of income inequality in achievement of model (19). Because certain variables such as family income have severe missing rates, it robustifies the MAR assumption to involve auxiliary covariates (e.g., parent occupation and pre-test score) correlated with missing values or patterns (Collins et al. 2003). We consider two approaches to augment auxiliary covariates, partially observed or measured with error, to the CD model. One approach is to assume such covariates to be linearly associated with the outcome and income. Violation of the linearity, however, may produce biased estimation. The other approach is to augment them as responses to $R^*_{ij}$, thereby allowing them to be non-linearly associated with the outcome and income. We then transform the MLE of the CD model to those of a nested model (19).

## 4.1 Linearly Associated Auxiliary Covariates

To augment auxiliary covariates that are linearly associated with the outcome and income, we extend the scalar $C^*_{ij}$ of model (12) to a vector $\boldsymbol{C}^*_{ij} = [C^*_{ij}\ A^{*T}_{1ij}\ A^{*T}_{2j}]^T$ consisting of income $C^*_{ij}$ and auxiliary covariates $A^*_{1ij}$ at level 1 and $A^*_{2j}$ at level 2. We then write the CD model

$$R^*_{ij} = (\gamma_{00} + u^*_{0j}) + (\gamma_{10} + u^*_{1j})\epsilon^*_{ij} + \gamma^T_{20}\epsilon^*_{1ij} + e^*_{ij}, \quad \boldsymbol{C}^*_{ij} = \boldsymbol{\delta} + \boldsymbol{\nu}_j + \boldsymbol{\varepsilon}^*_{ij} \tag{21}$$

where $\boldsymbol{\delta} = \left[\delta\ \delta^T_1\ \delta^T_2\right]^T$, $\boldsymbol{\nu}_j = \left[\nu_j\ \nu^T_{1j}\ \nu^T_{2j}\right]^T$ and $\boldsymbol{\varepsilon}^*_{ij} = \left[\epsilon^*_{ij}\ \epsilon^{*T}_{1ij}\ 0^T\right]^T = [\boldsymbol{\epsilon}^*_{ij}\ 0^T]^T$ for the means $[\delta^T_1\ \delta^T_2]^T$ and school-specific random effects $[\nu^T_{1j}\ \nu^T_{2j}]^T$ of $[A^{*T}_{1ij}\ A^{*T}_{2j}]^T$ and the child-specific random effects $\epsilon^*_{1ij}$ of $A^*_{1ij}$. Other components $\gamma_{00}$, $\gamma_{10}$, $u^*_{0j} = \gamma^T_{01}\boldsymbol{\nu}_j + u_{0j}$, $u^*_{1j} = \gamma^T_{11}\boldsymbol{\nu}_j + u_{1j}$ and $e^*_{ij}$ are defined in model (19) except the linear effects $\gamma_{20}$ of $\epsilon^*_{1ij}$. Again $e^*_{ij}$ and $\boldsymbol{\epsilon}^*_{ij} \sim N(0, \Sigma_\epsilon)$ are independent of each other and $u^*_j \sim N(0, \tau)$ for $\Sigma_\epsilon = \begin{bmatrix} \sigma_{cc} & \Sigma_{c1} \\ \Sigma_{1c} & \Sigma_{11} \end{bmatrix}$ and $u^*_j = [u^*_{0j}\ u^*_{1j}\ \boldsymbol{\nu}^T_j]^T$; let $cov(u^*_{0j}, \boldsymbol{\nu}_j) = \tau_{0\boldsymbol{\nu}}$, $cov(u^*_{1j}, \boldsymbol{\nu}_j) = \tau_{1\boldsymbol{\nu}}$ and $\boldsymbol{\nu}_j \sim N(0, \tau_{\boldsymbol{\nu}\boldsymbol{\nu}})$. Auxiliary predictors $(\nu_{1j}, \nu_{2j})$ and $\epsilon^*_{1ij}$ are linearly associated with the outcome and income at levels 2

and 1, respectively. The parameters are $\theta^*_{(21)} = (\gamma_{00}, \gamma_{10}, \gamma_{20}, \tau, \sigma^2, \boldsymbol{\delta}, \Sigma_\epsilon)$.

We select $u^*_{1j}$ provisionally known again. Let $Y^*_{ij} = [R^*_{ij}\ C^*_{ij}\ A^{*T}_{1ij}]^T$ and $A^*_{2j}$ be of respective lengths $p_1$ and $p_2$. Using model (13) now for $f(u^*_{0j}, \boldsymbol{\nu}_j|u^*_{1j})$ and $\alpha_{\boldsymbol{\nu}|1} = [\alpha_{\nu|1}\ \alpha^T_{\nu 1|1}\ \alpha^T_{\nu 2|1}]^T$, we find the provisional joint model (14) this time for $Y^*_j = [Y^{*T}_{1j} \cdots Y^{*T}_{n_j j}\ A^{*T}_{2j}]^T$ where

$$X^*_j = diag\{[X^{*T}_{11j} \cdots X^{*T}_{1n_j j}]^T, X^*_{2j}\}, \qquad \alpha = [\alpha^T_1\ \alpha^T_2]^T, \qquad Z^*_j = diag\{1_{n_j} \otimes I_{p_1}, I_{p_2}\},$$
$$b_j = [u^*_{0j} - \alpha_{0|1}u^*_{1j}\ \boldsymbol{\nu}^T_j - \alpha^T_{\boldsymbol{\nu}|1}u^*_{1j}]^T, \quad r^*_j = [r^{*T}_{1j} \cdots r^{*T}_{n_j j}\ 0^T]^T, \quad \Psi^*_j = diag\{I_{n_j} \otimes \Sigma^*_j, 0\}$$

for $X^*_{1ij} = [I_{p_1}\ I_{p_1} u^*_{1j}]$, $X^*_{2j} = [I_{p_2}\ I_{p_2} u^*_{1j}]$, $\alpha^T_1 = [\gamma_{00}\ \delta\ \delta^T_1\ \alpha_{0|1}\ \alpha_{\nu|1}\ \alpha^T_{\nu 1|1}]$, $\alpha^T_2 = [\delta^T_2\ \alpha^T_{\nu 2|1}]^T$,

$$r^*_{ij} = \begin{bmatrix} 1 & B^T_{1j} \\ 0 & I_{p_1-1} \end{bmatrix} \begin{bmatrix} e^*_{ij} \\ \boldsymbol{\epsilon}^*_{ij} \end{bmatrix}, \qquad \Sigma^*_j = \begin{bmatrix} B^T_{1j}\Sigma_\epsilon B_{1j} + \sigma^2 & B^T_{1j}\Sigma_\epsilon \\ \Sigma_\epsilon B_{1j} & \Sigma_\epsilon \end{bmatrix} \tag{22}$$

denoting $B^T_{1j} = [\gamma_{10} + u^*_{1j}\ \gamma^T_{20}]$.

We estimate $\theta^*_{(21)}$ using its one-to-one transformation $\theta = (\alpha, \Omega, \gamma_{10}, \gamma_{20}, \Sigma_\epsilon, \sigma^2, \tau_{11})$ of $h(Y^*_j|u^*_{1j})\phi(u^*_{1j}; 0, \tau_{11})$ by the EM algorithm, computing $h(Y_j)$ and $E(S_{CDj}|Y_j)$ by AGHQ as before. See Appendix B for the E step. The scalar PKRE $u^*_{1j}$ yields efficient computation.

To translate $\theta^*_{(21)}$ to $\theta^*_{(19)}$, we use model (13) to let $\beta_{kj} = \gamma_{k0} + u^*_{kj}$ and find

$$E(\beta_{kj}|\nu_j) = \gamma_{k0} + \gamma^*_{k1}\nu_j \text{ and } cov(\beta_{kj}, \beta_{k'j}|\nu_j) = \tau^*_{kk'} = \tau_{kk'} - \gamma^*_{k1}\gamma^*_{k'1}\tau_{\nu\nu}$$

for $\gamma^*_{k1} = \tau_{k\nu}/\tau_{\nu\nu}$ and $k, k' = 0, 1$. We then marginalize auxiliary $\epsilon^*_{1ij}$ out to obtain

$$E(R^*_{ij}|\epsilon^*_{ij}, \nu_j) = \gamma_{00} + \gamma^*_{01}\nu_j + (\gamma_{10} + \gamma^*_{11}\nu_j)\epsilon^*_{ij} + \gamma^T_{20}E(\epsilon^*_{1ij}|\epsilon^*_{ij}),$$
$$var(R^*_{ij}|\epsilon^*_{ij}, \nu_j) = \tau^*_{00} + 2\tau^*_{01}\epsilon^*_{ij} + \tau^*_{11}\epsilon^{*2}_{ij} + \gamma^T_{20}var(\epsilon^*_{1ij}|\epsilon^*_{ij})\gamma_{20} + \sigma^2$$

that should be of form $(\gamma_{00} + \gamma_{01}\nu_j) + (\gamma_{10} + \gamma_{11}\nu_j)\epsilon^*_{ij}$ and $\tau_{00|\nu} + 2\tau_{01|\nu}\epsilon^*_{ij} + \tau_{11|\nu}\epsilon^{*2}_{ij} + \sigma^2$, respectively. See Appendix D for detail. We illustrate this approach in sections 6 and 7.

### 4.2 Non-linearly Associated Auxiliary Covariates

The linearity assumption between $A^*_{1ij}$ and $(R^*_{ij}, C^*_{ij})$ in model (21) may be violated to produce biased estimation of $\theta^*_{(19)}$. In that case, we augment $A^*_{1ij}$ to multivariate $\boldsymbol{R}^*_{ij} = [R^*_{ij}\ A^{*T}_{1ij}]^T$ of length $r$, allowing them to be non-linearly associated in

$$\boldsymbol{R}^*_{ij} = (\gamma_{00} + u^*_{0j}) + (\gamma_{10} + u^*_{1j})\epsilon^*_{ij} + e^*_{ij}, \qquad e^*_{ij} \sim N(0, \Sigma_e) \tag{23}$$

and $\boldsymbol{C}^*_{ij} = [C^*_{ij}\ A^{*T}_{2j}]^T$ in Equations (21) for conformable vectors $\gamma_{00}$ and $\gamma_{10}$ of fixed effects and random vectors $u^*_{0j}$ and $u^*_{1j}$ independent of $e^*_{ij}$ as before. It is straightforward to find the provisional joint model (14) for $Y^*_{ij} = (\boldsymbol{R}^*_{ij}, C^*_{ij})$ and $A^*_{2j}$, selecting $r$-by-1 $u^*_{1j}$ to be provisionally known. We estimate the joint model efficiently and, then, translate the bivariate distribution of $R^*_{ij}$ and $C^*_{ij}$ to $\theta^*_{(19)}$; see Appendix D again for the translation. Numerical approximation is now intensive with respect to vector $u^*_{1j}$. Finally, some of $A^*_{1ij}$ may be linearly associated with the outcome and income while others may not. We illustrate this case in Data Analysis.

With additional known auxiliary covariates, we marginalize them out first given their expectation and covariance matrix estimated from sample before the translation above.

## 5 Within-Level Interactions and Polynomial Terms

We write a CD model including the level-2 interaction effects $\gamma_{04}$ of $\nu_j\nu_{1j}$

$$R^*_{ij} = (\gamma_{00} + \gamma_{01}\nu_j + \gamma^T_{02}\nu_{1j} + \gamma^T_{03}\nu_{2j} + \gamma^T_{04}\nu_j\nu_{1j} + u_{0j}) + \gamma^T_{10}\boldsymbol{\epsilon}^*_{ij} + e^*_{ij} \tag{24}$$

and $\boldsymbol{C}^*_{ij}$ as in model (21) where $u_{0j} \sim N(0, \tau_{00|\boldsymbol{\nu}})$ and $\boldsymbol{\epsilon}^*_{ij}$ has fixed effects $\gamma_{10}$ for simplicity. We select one interactive term $\nu_j$, $\leq$ the other in dimension, provisionally known to find

$$\begin{bmatrix} u_{0j} \\ \nu_{1j} \\ \nu_{2j} \end{bmatrix} \Bigg| \nu_j \sim N\left( \begin{bmatrix} 0 \\ \alpha_{\nu 1|\nu}\nu_j \\ \alpha_{\nu 2|\nu}\nu_j \end{bmatrix}, \Omega = \begin{bmatrix} \tau_{00|\boldsymbol{\nu}} & 0 & 0 \\ 0 & \tau_{\nu 1\nu 1|c} & \tau_{\nu 1\nu 2|c} \\ 0 & \tau_{\nu 2\nu 1|c} & \tau_{\nu 2\nu 2|c} \end{bmatrix} \right). \tag{25}$$

Let $b_j = [u_{0j} \;\; \nu_{1j}^T - \alpha_{\nu 1|\nu}^T \nu_j \;\; \nu_{2j}^T - \alpha_{\nu 2|\nu}^T \nu_j]^T = [u_{0j} \; b_{1j}^T \; b_{2j}^T]^T$ to find the model given $\nu_j$

$$\begin{aligned} R^*_{ij} &= X_{Rj}^T \alpha_R + Z_{Rj}^T b_j + \gamma_{10}^T \boldsymbol{\epsilon}^*_{1ij} + e^*_{ij} \sim N(X_{Rj}^T \alpha_R, Z_{Rj}^T \Omega Z_{Rj} + \gamma_{10}^T \Sigma_\epsilon \gamma_{10} + \sigma^2) \\ C^*_{ij} &= \delta + \nu_j + \epsilon^*_{ij}, \quad A^*_{1ij} = \delta_1 + \alpha_{\nu 1|\nu}\nu_j + b_{1j} + \epsilon^*_{1ij}, \quad A^*_{2j} = \delta_2 + \alpha_{\nu 2|\nu}\nu_j + b_{2j} \end{aligned} \tag{26}$$

for $X_{Rj}^T = [1 \; \nu_j \; \nu_j^2]$, $\alpha_R = \left[\gamma_{00} \; \gamma_{01} + \gamma_{02}^T \alpha_{\nu 1|\nu} + \gamma_{03}^T \alpha_{\nu 2|\nu} \; \gamma_{04}^T \alpha_{\nu 1|\nu}\right]^T$ and $Z_{Rj}^T = \left[1 \; \gamma_{02}^T + \gamma_{04}^T \nu_j \; \gamma_{03}^T\right]$ where $C^*_{ij}$ varies within but not between clusters. As in Section 4.1, we stack these equations to find the implied provisional model (14) for $Y^*_j = [Y^{*T}_{1j} \cdots Y^{*T}_{n_j} \; A^{*T}_{2j}]^T$, and compute $h(Y_j)$, setting $S_{CDj} = 1$ below, and $E(S_{CDj}|Y_j)$ by

$$h(Y_j)E(S_{CDj}|Y_j) = \int E(S_{CDj}|\nu_j, Y_j) h(Y_j|\nu_j) \phi(\nu_j; 0, \tau_{cc}) d\nu_j \tag{27}$$

numerically for $h(Y_j|\nu_j)$ from the provisional model.

We now consider another CD model including level-1 interaction effects

$$R^*_{ij} = (\gamma_{00} + \gamma_{01}^T \boldsymbol{\nu}_j + u_{0j}) + \gamma_{10}\epsilon^*_{ij} + (\gamma_{20}^T + \gamma_{30}^T \epsilon^*_{ij})\epsilon^*_{1ij} + e^*_{ij} \tag{28}$$

and $\boldsymbol{C}^*_{ij}$ in model (21) where $\epsilon^*_{ij}$ and $\epsilon^*_{1ij}$ have main ($\gamma_{10}$ and $\gamma_{20}$) and interaction effects ($\gamma_{30}$). We select an interactive term $\epsilon^*_{ij}$, $\leq \epsilon^*_{1ij}$ in dimension, provisionally known again, and find

$$\epsilon^*_{1ij}|\epsilon^*_{ij} \sim N(\alpha_{\epsilon 1|c}\epsilon^*_{ij}, \Sigma_{1|c}) \text{ for } \alpha_{\epsilon 1|c} = \Sigma_{1c}\sigma_{cc}^{-1} \text{ and } \Sigma_{1|c} = \Sigma_{11} - \alpha_{\epsilon 1|c}\sigma_{cc}\alpha_{\epsilon 1|c}^T.$$

Given $\epsilon^*_{ij}$ provisionally constant, let $u^*_{0j} = \gamma^T_{01}\boldsymbol{\nu}_j + u_{0j}$ and $a^*_{1ij} = \epsilon^*_{1ij} - \alpha_{\epsilon 1|c}\epsilon^*_{ij}$ to express

$$R^*_{ij} = X^T_{Rij}\alpha_R + u^*_{0j} + B^T_{1ij}a^*_{1ij} + e^*_{ij} \sim N(X^T_{Rij}\alpha_R, \tau_{00} + B^T_{1ij}\Sigma_{1|c}B_{1ij} + \sigma^2),$$
$$C^*_{ij}|\epsilon^*_{ij} \sim N(\delta + \epsilon^*_{ij}, \tau_{cc}), \quad A^*_{1ij} = \delta_1 + \alpha_{\epsilon 1|c}\epsilon^*_{ij} + \nu_{1j} + a^*_{1ij}, \quad A^*_{2j} = \delta_2 + b_{2j} \tag{29}$$

where $X^T_{Rij} = [1\ \epsilon^*_{ij}\ \epsilon^{*2}_{ij}]$, $\alpha_R = [\gamma_{00}\ \gamma_{10}{+}\gamma^T_{20}\alpha_{\epsilon 1|c}\ \gamma^T_{30}\alpha_{\epsilon 1|c}]$, $B^T_{1ij} = \gamma^T_{20}{+}\gamma^T_{30}\epsilon^*_{ij}$, $cov(R^*_{ij}, A^*_{1ij}|b_j, \epsilon^*_{ij}) = B^T_{1ij}\Sigma_{1|c}$ and $C^*_{ij}$ now varies between but not within clusters to imply the provisional model (14) given $\epsilon^*_j = (\epsilon^*_{1j}, \cdots, \epsilon^*_{n_jj})$ for $b_j = [u^*_{0j}\ \boldsymbol{\nu}^T_j]^T$. We compute $h(Y_j)$ and $E(S_{CDj}|Y_j)$ by

$$h(Y_j)E(S_{CDj}|Y_j) = \int E(S_{CDj}|\epsilon^*_j, Y_j)h(Y_j|\epsilon^*_j)\phi(\epsilon^*_j; 0, I_{n_j}\sigma_{cc})d\epsilon^*_j \tag{30}$$

numerically for $h(Y_j|\epsilon^*_j)$ from the provisional model as before. The numerical integral can be computationally intensive, in particular, given large cluster sizes; multivariate Laplace approximation (Pinheiro and Bates 1995; Raudenbush et al. 2000) and parallel computation of each cluster may result in efficient computation.

## 5.1 Rules for Choosing Provisionally Known Random Effects

We now provide general rules for selecting PKREs:

(i) For an interaction $\epsilon^*_{ij}\epsilon^*_{1ij}$ as in Equations (28), we hold $\epsilon^*_{ij}$, $\leq \epsilon^*_{1ij}$ in dimension, constant. The resulting model quadratic in $\epsilon^*_{ij}$ will minimize the dimension of AGHQ;

(ii) For a level-2 interaction $\nu_j\nu_{1j}$, we hold one with a smaller dimension constant again;

(iii) For a three-way interaction $\epsilon^*_{ij}\epsilon^*_{1ij}\epsilon^*_{2ij}$ at level 1, hold two terms $\leq$ the third one in dimension constant and this applies to a three-way interaction at level 2, too;

(iv) For cross-level interactions $\boldsymbol{\nu}_j\epsilon^*_{ij}$ as in model (21), hold $u^*_{1j} = \gamma^T_{11}\boldsymbol{\nu}_j + u_{1j}$ constant;

(v) For the cluster-specific effects $u^*_{1j}$ of $\epsilon^*_{ij}$, we hold $u^*_{1j}$ constant;

(vi) Finally, for a subset of these effects, hold constant the union of the PKREs of the subset.

Our CD model in each case includes a scientific model of interest and induces a provisional joint model (14). Because the models are one-to-one transformations of each other, the scientific model is guaranteed to be compatible with the joint model we estimate.

# 6 Data Analysis

Rising income inequality in the US and other nations has recently attracted substantial attention (Piketty, 2014). A key question involves the consequence of such inequality for equality of opportunity among children. Following past research, we decompose the association between family income and educational achievement into a contextual component and a child-specific component (Firebaugh, 1978; Willms, 1986; Lee and Bryk, 1989). The contextual component reflects the fact that elementary schools in the US are quite segregated based on family income. Such segregation reflects and may reinforce residential segregation as a function of family income. Two children having the same family income might differ in educational achievement as a result of their experience in low-income versus high-income schools. The individual component reflects socioeconomic inequality within schools. Children attending the same school who differ with respect to family income may tend to differ with respect to their achievement. However, the magnitude of this within-school disparity may vary from school to school (Raudenbush and Bryk, 1986; Lee and Bryk, 1989). The contextual effects model (Willms, 1986) supports the composition of inequality in achievement that we seek.

First, we decompose family income for child $i$ in school $j$ into between-school and within-school components as in model (19)

$$\ln(income_{ij}) = C^*_{ij} = \delta + \nu_j + \epsilon^*_{ij}, \;\; mathS_{ij} = R^*_{ij} = (\gamma_{00} + \gamma_{01}\nu_j + u_{0j}) + (\gamma_{10} + \gamma_{11}\nu_j + u_{1j})\epsilon^*_{ij} + e^*_{ij}$$

for the mean of log-income $\delta$, the school-specific deviation from the mean $\nu_j \sim N(0, \sigma_{cc})$, and the child-specific component $\epsilon^*_{ij} \sim N(0, \sigma_{cc})$. A child's mathematics achievement in spring 1999 (mathS) depends on these components via the model (19) for $mathS_{ij} = R^*_{ij}$. The parameters are $\theta^*_{(19)} = (\gamma_{00}, \gamma_{01}, \gamma_{10}, \gamma_{11}, \tau_{00|\nu}, \tau_{01|\nu}, \tau_{11|\nu}, \sigma^2, \delta, \tau_{\nu\nu}, \sigma_{cc})$.

We choose math achievement as our outcome because of its importance in predicting educational attainment and adult earnings (Nomi and Raudenbush, 2016; Rivera-Batiz, 1982). In model (19), $\gamma_{01}$ is the between-school gradient, reflecting the expected difference in $R^*_{ij}$ associated with a unit difference in school mean income; $\gamma_{10}$ reflects the average within-school gradient. However, the within-school gradient may depend on school mean income, an interaction effect represented by $\gamma_{11}$, and this gradient may also vary randomly over schools as represented by $u_{1j} \sim N(0, \tau_{11|\nu})$. School-mean achievement is $\gamma_{00}$ and, conditional on income, varies randomly over schools, $u_{0j} \sim N(0, \tau_{00|\nu})$. If the within-school gradients were constant ($\gamma_{11} = \tau_{11} = 0$), the overall linear coefficient for income will be

$$E(R^*_{ij}|C^*_{ij} = c + 1) - E(R^*_{ij}|C^*_{ij} = c) = \rho\gamma_{01} + (1 - \rho)\gamma_{10}$$

where $\rho = \tau_{cc}/(\tau_{cc} + \sigma_{cc})$ can be regarded as an index of school segregation as a function of income. Define $\gamma_c = \gamma_{01} - \gamma_{10}$ as the "contextual effect" (Willms, 1986), the expected difference in math achievement between two students with the same family income who attend two schools that differ by one unit in school mean income. A nation's income gradient would be $\rho\gamma_c + \gamma_{10}$, which increases with the within school segregation based on income, $\rho$, the contextual coefficient, $\gamma_c$ and within-school gradient $\gamma_{10}$. This simple relationship will not hold if the within-school gradients vary over schools, and one purpose of our analysis is to test that proposition.

Table 1: Each variable for analysis with mean (standard deviation (sd), missing %).

| symbol | name | mean (sd, missing %) | symbol | name | mean (sd, missing %) |
|---|---|---|---|---|---|
| $R^*_{ij}$ | mathS | 0.00 (1.00, 8) | $A^*_{1ij}$ | occupation | 0.00 (1.00, 5) |
| $C^*_{ij}$ | income | 10.73 ( 0.71, 32) | $A^*_{1ij}$ | age | 0.00 (1.00, 6) |
| $A^*_{1ij}$ | mathF | 0.00 (1.00, 12) | $A^*_{2j}$ | enrollment | 0.00 (1.00, 17) |

To do so, we use data from 21,211 children attending kindergarten in 1,018 schools as of fall 1998, a nationally representative sample known as the Early Childhood Longitudinal Study of 1998 ("ECLS") that is publicly available at https://nces.ed.gov/ecls; see Table 1. Only 8 percent of the math achievement data are missing. However, family income data are missing for 32% of the sample, a finding that is quite typical in surveys of educational achievement. Fortunately, ECLS (Tourangeau et al., 2009) provides data on auxiliary variables, including the maximum occupational status score of parents (occupation), missing for only 5% of the cases, as well as math achievement in fall 1998 (mathF), which is strongly predictive of math achievement in spring 1999 (mathS). In all, we have 4 auxiliary variables, correlated with income, outcome or missing patterns.

In this paper, we take convergence to ML to be less than $10^{-4}$ in the square root of the summed squared differences between $\hat{\theta}$ of two consecutive iterations. We estimate the model for covariates $\boldsymbol{C}_{ij}^*$ using all observed values by ML via the EM algorithm (SR 2007) and a scientific model for $R_{ij}^*$ given the covariates and their sample cluster and overall means by complete case analysis, and transform the estimates to the initial values $\hat{\theta}$ of the joint model. We carry out complete case analysis by R (R Core Team, 2017), estimate $\theta$ on a Dell XPS laptop with the 11th generation Intel(R) Core(TM) i9-11900H processor at 2.50GHz and 64 GB RAM, and test a hypothesis at a level $\alpha = 0.05$.

## 6.1 Linearly Associated Auxiliary Covariates

Recall from section 4 that we have two strategies. Following Section 4.1, we model the auxiliary covariates linearly associated with the outcome and income in model (21) where $A_{1ij}^*$ is a vector of mathF, occupation, and age in months at assessment of spring 1999 (age) and $A_{2j}^*$ is the square root of kindergarten enrollment (enrollment) by the Box-Cox transformation. Consequently, $\theta_{(21)}^* = (\gamma_{00}, \gamma_{10}, \gamma_{20}, \tau, \sigma^2, \boldsymbol{\delta}, \Sigma_\epsilon)$ consists of 10 fixed effects (3-by-1 $\gamma_{20}$ and 5-by-1 $\boldsymbol{\delta}$) and 39 variances and covariances (7-by-7 $\tau$ and 4-by-4 $\Sigma_\epsilon$).

We standardized each variable to have mean 0 and variance 1, except income for inter-

pretation. Estimation of $\theta^*_{(21)}$ with a provisionally known $u^*_{1j}$ and 20 abscissas converged fast to ML in 9 iterations and 13 seconds. The transformed estimates $\hat{\theta}^*_{(19)}$ and standard errors (SEs), multiplied by 100, are listed under "EM-AGHQ I" in Table 2. The school-mean and within-school components of family income are positively associated with math achievement while their interaction effect is insignificant. The within-school income effects appear to vary at most modestly across schools with the variance estimate $\hat{\tau}_{11|\nu} = 0.19$ less than the associated SE 0.25. Based on $\ln(\hat{\tau}_{11|\nu}) \sim N[\ln(\tau_{11|\nu}), var(\hat{\tau}_{11|\nu})/\tau^2_{11|\nu}]$, we find a large-sample 95% confidence interval (CI) for $\tau_{11|\nu}$: $(0.01, 2.62)$ near zero.

Table 2: Estimates×100 (standard errors×100) of model (19) by EM-AGHQ I and II.

| predictor | parameter | EM-AGHQ I | EM-AGHQ II |
|---|---|---|---|
| 1 | $\gamma_{00}$ | 0.36 (2.38) | 0.20 (2.46) |
| $\nu_j$ | $\gamma_{01}$ | 78.69 (2.93)** | 77.22 (3.01)** |
| $\epsilon^*_{ij}$ | $\gamma_{10}$ | 28.06 (1.36)** | 37.38 (1.74)** |
| $\epsilon^*_{ij}\nu_j$ | $\gamma_{11}$ | 0.77 (2.03) | 14.33 (3.58)** |
| | $\tau_{00\|\nu}$ | 9.30 | 9.64 |
| | $\tau_{01\|\nu}$ | 0.64 | 2.35 |
| | $\tau_{11\|\nu}$ | 0.19 (0.25) | 3.08 (0.69)** |
| | 95% CI for $\tau_{11\|\nu}$ | (0.01, 2.62) | (1.98, 4.78) |
| | $\sigma^2$ | 75.21 | 73.74 |
| | log joint ML | -107361.00 | -107177.03 |

*: p-value<.05; **: p-value<.01

To test this model against the null hypothesis that $\gamma_{11} = 0$ and $\tau_{11|\nu} = 0$, we estimated the null model, the multivariate normal distribution of linearly associated $(R^*_{ij}, C^*_{ij}, A^*_{1ij}, A^*_{2j})$, efficiently by the EM algorithm (SR 2007, 2010). The model consisted of 42 parameters comprising 6 fixed intercepts and 6-by-6 level-2 and 5-by-5 level-1 variance covariance matrices, and converged to log ML -107370.60. Compared to the log ML of $\theta^*_{(21)}$ displayed at the last row of Table 2, the likelihood ratio test statistic to test the two joint models is 19.20 with 7 degrees of freedom to give a conservative p-value $< 0.01$ (Stram and Lee 1994). Therefore, we infer that the outcome is non-linearly associated with income. Lastly, estimation using 10 abscissas also produced the same estimates under EM-AGHQ I.

## 6.2 Non-linearly Associated Auxiliary Covariates

Preliminary analysis indicates that mathF and occupation may be nonlinearly associated with the outcome and income. To test the hypothesis, following Section 4.2, we model multivariate responses $\boldsymbol{R}^*_{ij}$ =(mathS, mathF, occupation) and $A^*_{1ij}$ =age in model (21):

$$\boldsymbol{R}^*_{ij} \;=\; (\gamma_{00} + u^*_{0j}) + (\gamma_{10} + u^*_{1j})\epsilon^*_{ij} + \gamma_{20}\epsilon^*_{1ij} + e^*_{ij} \tag{31}$$

and $\boldsymbol{C}^*_{ij} = [C_{ij}\ A^*_{1ij}\ A^*_{2j}]^T$ as before for 3-by-1 vectors $\gamma_{00}$, $\gamma_{10}$ and $\gamma_{20}$ of fixed effects and random vectors $u^*_{0j}$, $u^*_{1j}$ and $e^*_{ij} \sim N(0, \Sigma_e)$. Therefore, $\tau$ is 9-by-9, $\Sigma_e$ 3-by-3 and $\Sigma_\epsilon$ 2-by-2; the CD model comprises 12 fixed effects and 54 variances and covariances. With 3-by-1 $u^*_{1j}$ provisionally known, we estimated the joint model using 10 abscissae per dimension which converged to ML in 734th iterations and 928 minutes. The log ML of $\theta^*_{(31)}$ is shown at the bottom row under EM-AGHQ II. The LRT statistic to test $H_0$ : model (21) vs $H_1$ : model (31) is 367.93. The conservative LRT with 17 degrees of freedom (Stram and Lee 1994) produces a p-value 0 to reject the null in favor of mathF or occupation nonlinearly associated with the outcome and income.

The translated estimates $\hat{\theta}^*_{(19)}$ and SEs are listed under EM-AGHQ II in Table 2. Compared to those under EM-AGHQ I, the main effect of within-school income is larger; furthermore, the interaction effect is significant, and so is the random effects of income by the Wald test to produce a 95% CI for $\tau_{11}$ now distant from zero. We conclude that the linearity assumption associated with Equations (21) is violated to attenuate the main, interaction and random effects of within-school income $\epsilon^*_{ij}$, confounded with the auxiliary covariates. As a result, EM-AGHQ II produces a smaller $\hat{\sigma}^2 = 73.74$ and, thus, explains more outcome variability within schools than does EM-AGHQ I.

### 6.3 Known Auxiliary Covariates

Either model (21) or (31) may be extended to control for known auxiliary covariates such as race ethnicity and gender at level 1 and school location and sector at level 2. The joint model may be estimated given the provisionally known $u_{1j}^*$ again, and translated to $\hat{\theta}_{(19)}^*$ and SEs. See Appendices B and D for detail.

## 7 Simulation Study

We focus on ML estimation of the scientific HLM (19) after simulating outcome $R_{ij}^*$ and income $C_{ij}^*$ from a joint model conditional on auxiliary covariates within which the HLM is nested. The goal is to compare our estimators (EM-AGHQ) with those by four methods: 1) the benchmark method (BM) given $\nu_j$ and $\epsilon_{ij}^*$; 2) complete-case analysis (CC) given $\bar{C}_j^* - \bar{C}^*$ and $C_{ij}^* - \bar{C}_j^*$ instead; 3) MLE on MI (SR 2007, 2010); and 4) the Gibbs sampler (GS) of Enders et al. (2020) implemented in software Blimp (Keller and Enders 2021). BM is based on complete data while others are based on data MAR. Therefore, a good method will produce estimates near the BM counterparts. BM and CC estimate the scientific model by the lme4 package (Bates et al. 2015) in R. MLE on MI uses C programs to estimate incompatible MHLM (8) by ML and impute missing values including latent school mean incomes 20 times, more than did past multilevel missing data analyses (Schafer and Yucel 2002; SR 2007, 2013; Enders et al. 2020), from their predictive distribution given observed data implied by the MHLM at ML (SR 2007, 2010); and estimates the HLM given the MI by lme4. GS estimates the joint model, simultaneously generating 20 imputations of missing values excluding latent school mean incomes, by Blimp and, then, the HLM (19) given the MI by Blimp again. EM-AGHQ estimates the joint model by our C program and translates the estimates to the desired MLE in R.

We simulate the ECLS $R_{ij}^*$ and $C_{ij}^*$ closely in terms of sample sizes, correlations and missing rates. Specifically, for $n = 20$ children in each of $J = 1000$ schools, we simulate:

i) known auxiliary covariates $X_{21j} \sim Bernoulli(0.3)$ and $X_{1ij} \sim Bernoulli(0.45)$ equal to 1 (0) for a private (public) school and a minority (white) child, respectively, where $X_{1ij}$ varies within, but not between, schools for simplicity; ii) random intercepts and slope $\beta_{0j} = \gamma_{00}^T X_{2j} + u_{0j}^*$, $\beta_{1j} = \gamma_{10}^T X_{2j} + u_{1j}^*$ and $\beta_{Cj} = \delta_{00}^T X_{2j} + \nu_j$ from $N(1 + X_{21j}, 1)$ with covariances 0.8 for $X_{2j} = [1\ X_{21j}]^T$; iii) independent $e_{ij}^*, \epsilon_{ij}^* \sim N(0, 10)$ to simulate the CD joint model

$$R_{ij}^* = \beta_{0j} + \beta_{1j}(C_{ij}^* - \beta_{Cj}) + \gamma_{20} X_{1ij} + e_{ij}^*, \qquad C_{ij}^* = \beta_{Cj} + \delta_{10} X_{1ij} + \epsilon_{ij}^*. \tag{32}$$

The simulated parameters in $\theta$ consist of $\gamma_{00}^T = \gamma_{10}^T = \delta_{00}^T = [1\ 1]$, $\gamma_{20} = \delta_{10} = 1$, variances $\tau_{00} = \tau_{11} = \tau_{\nu\nu} = 1$ and $\sigma_{cc} = \sigma^2 = 10$, and covariances $\tau_{01} = \tau_{0\nu} = \tau_{1\nu} = 0.8$. We marginalize $X_{1ij}$ and $X_{21j}$ out given their simulated expectations and variances, and translate $\theta$ to the scientific model in column two of Table 3 as explained in Appendix D.

Table 3: Scientific model (19) estimated by BM, CC, MLE on MI, GS and EM-AGHQ. Each estimate or cell occupies two rows: % bias (average estimated SE) in the first row and the empirical estimate of true SE over simulated samples and a 95% coverage probability of the estimator in the next.

| covariate | simulated | BM | CC | MLE on MI | GS | EM-AGHQ |
|---|---|---|---|---|---|---|
| 1 | $\gamma_{00}$=2.34 | .06% (.03) | -8.32% (.05) | -8.16% (.05) | -.37% (.07) | .04% (.06) |
| | | .03, .96 | .08, .11 | .06, .09 | .07, .94 | .07, .92 |
| $\nu_j$ | $\gamma_{01}$=1.21 | .12% (.03) | .29% (.03) | -5.81% (.06) | -.59% (.05) | .07% (.05) |
| | | .03, .95 | .05, .85 | .06, .76 | .05, .94 | .05, .94 |
| $\epsilon_{ij}^*$ | $\gamma_{10}$=1.32 | .10% (.02) | .16% (.03) | -5.14% (.03) | -.26% (.04) | .06% (.03) |
| | | .02, .95 | .04, .85 | .03, .33 | .04, .95 | .04, .93 |
| $\epsilon_{ij}^*\nu_j$ | $\gamma_{11}$=0.83 | -.07% (.02) | -42.05% (.02) | -54.00% (.03) | -.52% (.03) | -.04% (.03) |
| | | .02, .96 | .02, .00 | .02, .00 | .03, .95 | .03, .95 |
| | $\sqrt{\tau_{00\|\nu}}$=0.77 | .00% ( - ) | 60.31% ( - ) | 22.83% ( - ) | .36% ( - ) | .09% (.05) |
| | | .03, - | .06, - | .05, - | .05, - | .05, .96 |
| | $\tau_{01\|\nu}$=0.33 | -.05% ( - ) | -28.75% ( - ) | -34.68% ( - ) | -.64% ( - ) | -.29% (.04) |
| | | .02, - | .06, - | .03, - | .04, - | .04, .95 |
| | $\sqrt{\tau_{11\|\nu}}$=0.61 | -.13% ( - ) | 39.00% ( - ) | 7.31% ( - ) | 2.69% ( - ) | -.27% (.03) |
| | | .01, - | .02, - | .02, - | .03, - | .03, .95 |
| | $\sqrt{\sigma^2}$=3.20 | -.02% ( - ) | .07% ( - ) | 25.39% ( - ) | .02% ( - ) | -.05% (.02) |
| | | .02, - | .02, - | .05, - | .02, - | .02, .96 |

Next, we simulate the ECLS missing rates closely by

$$logit(p_{ij}) \;\; = \;\; \phi_1 X_{1ij} + \phi_2(1 - X_{1ij}) + z_j, \;\; z_j \sim N(0,1) \tag{33}$$

given the known level-1 covariate $X_{1ij}$: missing values drawn from $Bernoulli(p_{ij})$ are MAR. Because the mechanism does not provide information about the simulated model (32), the parameter spaces of the missing data mechanism and joint model are also distinct. We simulate higher missing rates for minority than white students by $\phi_1 > \phi_2$: $\phi_1 = -0.2 > \phi_2 = -1.2$ for income $C^*_{ij}$ with a 35% missing rate (46% for $X_{1ij} = 1$, 27% for $X_{1ij} = 0$); and $\phi_1 = -2 > \phi_2 = -3$ for response $R^*_{ij}$ with an 11% missing rate (16% for $X_{1ij} = 1$, 7% for $X_{1ij} = 0$) on average.

We repeated simulating data and estimating the scientific model by the approaches 500 times to compute the % bias, average estimated SE (ASE), empirical estimate of the true SE (ESE) over samples and coverage probability (coverage) of each estimator in the next five columns. Each cell or estimate occupies two rows: % bias (ASE) in the first, and ESE and coverage in the next row. The lme4 package is unable to produce ESE and coverage of a variance or covariance estimate. The BM estimates are of course very accurate and precise with $\leq 0.13\%$ bias, small ASE close to ESE, and good coverages near the nominal 0.95 in column three.

The CC estimates in column four, however, are biased despite the large sample sizes. The standard deviations (SDs) $\sqrt{\tau_{00|\nu}}$ and $\sqrt{\tau_{11|\nu}}$ are 60% and 39% biased upward while the intercept $\gamma_{00}$, interaction effect $\gamma_{11}$ and covariance $\tau_{01|\nu}$ are 8%, 42% and 29% biased downward, respectively. Only the estimates of $\gamma_{01}$, $\gamma_{10}$ and $\sigma^2$ are comparable in accuracy to those by BM as Equations (34) reveal in Appendix A. The coverages are low with a zero coverage for $\gamma_{11}$. Finally, the uncertainty associated with the estimator of a cluster-level effect $\gamma_{01}$ seems underestimated by ASE smaller than ESE.

MLE on MI generates incompatible MI based on MHLM (8) without consideration of the

interaction effect and PKRE, thereby producing all estimates biased in column five that do not seem better than the CC estimates. Fixed effects and covariance are biased downward, and SDs upward. SEs are close to but coverages lower than CC counterparts.

On the contrary, GS without consideration of a PKRE produces estimates in column six nearly as accurate as BM counterparts except the SD of the random slope that is biased upward by 2.69% while EM-AGHQ yields all estimates in the last column as accurate or almost as accurate as BM estimates. Overall, both approaches produce estimates slightly less precise than BM estimates; ASE and ESE appear slightly larger than those by BM to reflect extra uncertainty due to latent covariates and missing values.

**Computation.** We used 20 abscissas to estimate the joint model by EM-AGHQ. Given data MAR, the estimation converged 450 times taking 101.6 iterations on average and 189 iterations at maximum, but did not converge until and was stopped to produce the estimates at the 300th iteration 50 times (10%). This does not appear to be the weakness of our approach as the convergence issue also occurred to each of BM and CC estimations producing 50 or more warnings of a model failing to converge. The convergence issue seems partly due to high missing rates but few covariates to explain missing values and patterns. In our experience thus far, the convergence rates seem positively associated with more covariates or abscissas. For example, this simulation using 10 abscissas resulted in practically identical estimates, but lowered the convergence rate given data MAR.

Blimp estimates 21 models per simulated data set: joint model (32) and the HLM given each of 20 imputations. We set 20000 burn-in and 10000 post burn-in iterations to estimate the joint model and impute MI, and 5000 burn-in and 5000 post burn-in iterations to estimate the HLM given the MI. These settings are based on preliminary analysis of five simulated data sets that produced the potential scale reduction statistics of all estimates of each model lower than or near 1.1 to imply a reasonable convergence to posterior distributions (Gelman and Rubin 1992).

# 8 Discussion

In this paper, we have considered how to estimate a two-level hierarchical linear model (HLM) efficiently where a continuous response $R^*$ and continuous covariates $C^*$ may be MAR and $C^*$ may have interactive, polynomial or randomly varying effects. Non-linearities of $C^*$ imply a nonstandard joint model $h(R^*, C^*) = h(Y^*)$ where $Y^* = (Y, Y_{mis})$ for observed $Y$ and missing $Y_{mis}$. The key idea is to introduce a unique factorization of the joint model involving "provisionally known" random effects (PKREs) $u$ such that the observed joint model $h(Y|u) = \int h(Y|\nu, u)g(\nu|u)d\nu$ is an analytically tractable multivariate normal (MN) theory HLM with respect to a high-dimensional random vector $\nu$. We computed the likelihood $h(Y) = \int h(Y|u)g(u)du$ numerically with respect to a low dimensional $u$ by means of adaptive Gauss-Hermite quadrature (AGHQ). The HLM involved random effects as predictors, reducing bias due to measurement error. The joint model $h(Y^*|u)g(u)$ induced by the HLM is guaranteed to be compatible with the HLM. We suggested general rules for selecting the PKREs in a way that minimized the dimension of AGHQ. Although useful for the HLMs considered in this paper, they are yet to be extended to other models, for example, for discrete outcomes. We hope that our work will spur research on estimation via PKREs.

The non-linearities of multiple covariates, multiple outcomes and/or the presence of partially observed discrete variables will increase the dimension of PKREs and, thus, the expense of numerical integration by AGHQ. In that case, integration via multivariate Laplace approximation may contribute to efficient computation (Pinherio and Bates 1995; Raudenbush et al. 2000).

Further research may address the problem of highly correlated random effects at the cluster level. One strategy would introduce shared random effects to cope with the "curse of dimensionality" by AGHQ as well as the multicollinearity (Miyazaki and Frank 2006; Sun et al. 2023). In addition, parallel computation of numerical integrals for groups of or single clusters will reduce per-iteration computation time while application of the parameter-extended EM algorithm (Liu et al. 1998) may reduce the number of iterations to converge.

In related research, Rockwood (2020) estimated a multilevel structural equations model by ML, integrating linear random effects conditional on nonlinear random effects analytically and, then, nonlinear effects numerically by Gaussian quadrature. Our analysis encountered both outcome and predictors quite severely missing. To simulate the analysis closely, we simulated a MAR mechanism due to a known auxiliary predictor. We leave the important extension to a MAR (Grund et al. 2021) or other mechanism due to the fully observed or missing values of the outcome to near future.

It is possible to extend and automate our program that enables a user to specify an analytic HLM and determines PKREs based on a set of rules given the HLM. To that end, we need to develop a more general set of rules, for example, involving discrete covariates MAR.

Often, MI of a binary predictor MAR under multivariate normality is efficiently analyzed (Schafer 1997; Grund et al, 2018). The MI will be, however, incompatible with a HLM having the nonlinear effect of the predictor and, thus, unable to always guarantee unbiased estimation of the HLM. We are currently extending our ML approach via the PKRE idea that will ensure compatibility with and, thus, produce unbiased estimation of a HLM having the nonlinear effects of categorical predictors.

Extension of our approach to MI via Bayesian methods may increase the robustness of findings and is straightforward. In particular, our MN joint model $h(Y^*, \nu|u;\theta) = h(Y_{mis}, Y, \nu|u;\theta)$ given the PKRE $u$ implies estimation of $\theta$ by the Gibbs sampler. The sampler will impute $(Y_{mis}, \nu, u, \theta)$ from their posteriors compatible with the joint model $h(Y^*, \nu|u, \theta)g(u|\theta)p(\theta)$ for a reasonably assumed prior $p(\theta)$ by drawing: i) $Y_{mis}$ and $\nu$ from MN $h(Y_{mis}, \nu|Y, u, \theta)$; ii) $u$ from nonstandard $g(u|Y^*, \nu, \theta) = h(Y^*, \nu|u, \theta)g(u|\theta)/h(Y^*, \nu|\theta)$, for example, by importance sampling via Markov Chain Monte Carlo integration of $h(Y^*, \nu|\theta) = E[h(Y^*, \nu|u, \theta)]$ that samples $u$ from a normal prior $g(u|\theta)$; and iii) $\theta$ from a standard posterior $p(\theta|Y^*, u, \nu)$ (Schafer and Yucel 2002). A potential virtue of a PKRE is to minimize the dimension of sampling the PKRE from a nonstandard posterior by importance sampling.

We find it important to solve the measurement error problem by including level-2 random effects as latent covariates. To that end, we also explain how the Gibbs sampler without consideration of a PKRE (Goldstein et al. 2014; Enders et al. 2020) may be modified to be compatible with our scientific model conditional on a latent covariate $\nu$ in Appendix E. A valuable future study is to compare the proposed Gibbs sampler estimators with existing estimators of a more sophisticated HLM, for example, involving multiple nonlinear effects or outcomes.


# Acknowledgment

We thank two anonymous reviewers and an associate editor for their helpful comments, Craig Enders and Brian Keller for providing Blimp simulation codes, and Dongho Shin for helping Blimp simulation in R environment. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D210022. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. The authors report there are no competing interests to declare.


## Appendix A: Problem of Bias in Estimating Model (19)

Let $R^*_{ij} = R_{ij}$ and $C^*_{ij} = C_{ij}$ fully observed, $\delta = 0$ to simplify notation, $\beta_{kj} = \gamma_{k0} + u^*_{kj} \sim N(\gamma_{k0}, \tau_{kk})$ and $cov(u^*_{0j}, u^*_{1j}) = \tau_{01}$ for $k = 0, 1$ in model (19). Because $\bar{C}_{\cdot j}$ and $\beta_j = (\beta_{0j}, \beta_{1j}, \nu_j)$ are independent of $C_{ij} - \bar{C}_{\cdot j}$, and $\epsilon^*_{ij}|C_{ij} - \bar{C}_{\cdot j} \sim N(C_{ij} - \bar{C}_{\cdot j}, \sigma_{cc}/n_j)$,

$$\begin{bmatrix} R_{ij} \\ \bar{C}_{\cdot j} \end{bmatrix} \Big| \beta_j, C_{ij} - \bar{C}_{\cdot j} \sim N\left( \begin{bmatrix} \beta_{0j} + \beta_{1j}(C_{ij} - \bar{C}_{\cdot j}) \\ \nu_j \end{bmatrix}, \begin{bmatrix} \beta^2_{1j}\sigma_{cc}/n_j + \sigma^2 & \beta_{1j}\sigma_{cc}/n_j \\ \beta_{1j}\sigma_{cc}/n_j & \sigma_{cc}/n_j \end{bmatrix} \right)$$

implying a mixed model $R_{ij}|C_{ij} - \bar{C}_{\cdot j} \sim N\left[\gamma_{00} + \gamma_{10}(C_{ij} - \bar{C}_{\cdot j}), var(R_{ij}|C_{ij} - \bar{C}_{\cdot j})\right]$ for

$$var(R_{ij}|C_{ij} - \bar{C}_{\cdot j}) = \tau_{00} + 2\tau_{01}(C_{ij} - \bar{C}_{\cdot j}) + \tau_{11}(C_{ij} - \bar{C}_{\cdot j})^2 + (\tau_{11} + \gamma^2_{10})\sigma_{cc}/n_j + \sigma^2,$$
$$cov(R_{ij}, \bar{C}_{\cdot j}|C_{ij} - \bar{C}_{\cdot j}) = \gamma_{01}\tau_{\nu\nu} + \gamma_{11}\tau_{\nu\nu}(C_{ij} - \bar{C}_{\cdot j}) + \gamma_{10}\sigma_{cc}/n_j.$$

Let $\lambda_j = \tau_{\nu\nu}/(\tau_{\nu\nu} + \sigma_{cc}/n_j)$ be the reliability of $\bar{C}_{\cdot j}$ as an error-prone measure of $\nu_j$ (Raudenbush and Bryk 2002). The implied $R_{ij}|C_{ij} - \bar{C}_{\cdot j}, \bar{C}_{\cdot j} \sim N(\mu_{ij}, V_{ij})$ has

$$\begin{aligned}
\mu_{ij} &= \gamma_{00} + [\gamma_{01} - (1-\lambda_j)(\gamma_{01} - \gamma_{10})]\,\bar{C}_{\cdot j} + \gamma_{10}(C_{ij} - \bar{C}_{\cdot j}) + \lambda_j\gamma_{11}\bar{C}_{\cdot j}(C_{ij} - \bar{C}_{\cdot j}) \\
V_{ij} &= \sigma^2 + \left[\tau_{00|\nu} + (1-\lambda_j)(\gamma_{01} - \gamma_{10})^2\tau_{\nu\nu} + (\tau_{11|\nu} + \gamma_{11}^2\tau_{\nu\nu})\sigma_{cc}/n_j\right] \qquad (34) \\
&+ 2\left[\tau_{01|\nu} + (1-\lambda_j)(\gamma_{01} - \gamma_{10})\gamma_{11}\tau_{\nu\nu}\right](C_{ij} - \bar{C}_{\cdot j}) + \left[\tau_{11|\nu} + (1-\lambda_j)\gamma_{11}^2\tau_{\nu\nu}\right](C_{ij} - \bar{C}_{\cdot j})^2.
\end{aligned}$$

The bias terms are complicated functions of cluster sizes $n_j$ and parameters, but revealing in the balanced case of $n_j = n$ where $\lambda_j = \lambda$. The interaction effect $\lambda\gamma_{11}$ of $\bar{C}_{\cdot j}(C_{ij} - \bar{C}_{\cdot j})$ has a downward bias term $-(1-\lambda)\gamma_{11}$ that introduces bias $(1-\lambda)(\gamma_{01} - \gamma_{10})\gamma_{11}\tau_{\nu\nu}$ and $(1-\lambda)\gamma_{11}^2\tau_{\nu\nu}$ in estimation of $\tau_{01|\nu}$ and $\tau_{11|\nu}$, respectively. Likewise, the main effect of $\bar{C}_{\cdot j}$ has a bias term $-(1-\lambda)(\gamma_{01} - \gamma_{10})$ which propagates bias $(1-\lambda_j)(\gamma_{01} - \gamma_{10})^2\tau_{\nu\nu}$ and $(1-\lambda)(\gamma_{01} - \gamma_{10})\gamma_{11}\tau_{\nu\nu}$ in estimation of $\tau_{00|\nu}$ and $\tau_{01|\nu}$, respectively. Estimation of $\tau_{00|\nu}$ results in an additional upward bias term $(\tau_{11|\nu} + \gamma_{11}^2\tau_{\nu\nu})\sigma_{cc}/n$ from the error-prone measure $\bar{C}_{\cdot j}$ of $\nu_j$. Consequently, this approach results in biased estimation of $(\gamma_{01}, \gamma_{11}, \tau_{00|\nu}, \tau_{01|\nu}, \tau_{11|\nu})$. In particular, the estimate of $\gamma_{11}$ is biased downward, but those of $\tau_{00|\nu}$ and $\tau_{11|\nu}$ upward.

Two special cases are of interest. When $\gamma_{11} = 0$, $cov(\beta_{0j}, \beta_{1j}|\bar{C}_{\cdot j}) = \tau_{01|\nu}$ and $var(\beta_{1j}|\bar{C}_{\cdot j}) = \tau_{11|\nu}$ become unbiased, and the estimator of $\tau_{00|\nu}$ becomes less biased. As cluster sizes $n_j \to \infty$, $\lambda_j \to 1$, $\bar{C}_{\cdot j} \to \nu_j$ by the laws of large numbers, and all bias terms tend to zero.

## Appendix B: The E Step for estimation of model (14)

Let $\boldsymbol{A}^*_{1ij} = [C^*_{ij}\ A^{*T}_{1ij}]^T$, $Y^*_{ij} = [R^*_{ij}\ \boldsymbol{A}^{*T}_{1ij}]^T$ $p_1$-by-1, $A^*_{2j}$ $p_2$-by-1, and $\boldsymbol{\nu}_{1j} = [\nu_j\ \nu_{1j}^T]^T$. A reasonably general CD model given known covariates $X_{1ij}$ at level 1 and $X_{2j}$ at level 2 is

$$\begin{aligned}
R^*_{ij} &= \gamma_{00}^T X_{2j} + u^*_{0j} + B_{1j}^T(\boldsymbol{A}^*_{1ij} - \Delta_{00}X_{2j} - \boldsymbol{\nu}_{1j}) + \gamma_{30}^T X_{1ij} + e^*_{ij} \\
\boldsymbol{A}^*_{1ij} &= \Delta_{00}X_{2j} + \Delta_{10}X_{1ij} + \boldsymbol{\nu}_{1j} + \boldsymbol{\epsilon}^*_{ij}, \quad A^*_{2j} = \Delta_2 X_{2j} + \nu_{2j} \qquad (35)
\end{aligned}$$

for $B_{1j}^T = [\gamma_{10}^T X_{2j} + u_{1j}^* \ \gamma_{20}^T]$, $\Delta_{00} = [\delta_{001} \ \Delta_{002}^T]^T$ and $\Delta_{10} = [\delta_{101} \ \Delta_{102}^T]^T$. Denote $\nu_{0j} = [u_{0j}^* \ \boldsymbol{\nu}_{1j}^T]^T$ to separate all other random effects $\nu_j^* = [\nu_{0j}^T \ \nu_{2j}^T]^T$ from $u_{1j}^*$ at level 2, and let

$$var(\nu_j^*) = \begin{bmatrix} T_{00} & T_{02} \\ T_{20} & T_{22} \end{bmatrix}, \ cov(\nu_j^*, u_{1j}^*) = \begin{bmatrix} T_{01} \\ T_{21} \end{bmatrix} \text{ and } var(\nu_j^* | u_{1j}^*) = \begin{bmatrix} T_{00|1} & T_{02|1} \\ T_{20|1} & T_{22|1} \end{bmatrix}.$$

Define matrix $O_{2j}$ that selects observed values $A_{2j} = O_{2j} A_{2j}^*$ in $A_{2j}^*$ such that $var(A_{2j}|u_{1j}^*) = O_{2j} T_{22|1} O_{2j}^T = T_{22|1j}$, and $cov(\nu_{aj}, \nu_{bj} | u_{1j}^*, A_{2j}) = \Omega_{abj} = T_{ab|1} - T_{a2|1} O_{2j}^T T_{22|1j}^{-1} O_{2j} T_{2b|1}$ for $a, b = 0, 2$. The likelihood $L(\theta) = \prod_j \int h(Y_j|u_{1j}^*) \phi(u_{1j}^*|0, \tau_{11}) du_{1j}^*$ has a key component

$$\begin{aligned} h(Y_j|u_{1j}^*) \propto & \left( |\Omega_{00j}|^{-1} |\Delta_j|^{-1} |T_{22|1j}|^{-1} \prod_i |\Sigma_{ij}|^{-1} \right)^{1/2} exp \left\{ -\frac{1}{2} \right. \\ & \left[ \sum_i e_{o1ij}^T \Sigma_{ij}^{-1} e_{o1ij} - \sum_i e_{o1ij}^T \Sigma_{ij}^{-1} O_{ij} \Delta_j^{-1} \left( \sum_i O_{ij}^T \Sigma_{ij}^{-1} e_{o1ij} + 2\Omega_{00j}^{-1} T_{02|1} O_{2j}^T T_{22|1j}^{-1} e_{o2j} \right) \right. \\ & \left. \left. + \ e_{o2j}^T \left( T_{22|1j}^{-1} O_{2j} T_{20|1} (\Omega_{00j}^{-1} - \Omega_{00j}^{-1} \Delta_j^{-1} \Omega_{00j}^{-1}) T_{02|1} O_{2j}^T T_{22|1j}^{-1} + T_{22|1j}^{-1} \right) e_{o2j} \right] \right\} \end{aligned} \tag{36}$$

for $e_{o1ij} = O_{ij}(d_{ij}^* - T_{01} \tau_{11}^{-1} u_{1j}^*)$, $e_{o2j} = O_{2j}(A_{2j}^* - \Delta_2 X_{2j} - T_{21} \tau_{11}^{-1} u_{1j}^*)$ and $\Delta_j = \sum_{i=1}^{n_j} A_{OOij} + \Omega_{00j}^{-1}$ where $A_{OOij} = O_{ij}^T \Sigma_{ij}^{-1} O_{ij}$ and $d_{ij}^* = \begin{bmatrix} R_{ij}^* \\ \boldsymbol{A}_{1ij} \end{bmatrix} - \begin{bmatrix} \gamma_{00}^T \\ \Delta_{00} \end{bmatrix} X_{2j} - \begin{bmatrix} 1 & B_{1j}^T \\ 0 & I_{p_1 - 1} \end{bmatrix} \begin{bmatrix} \gamma_{30}^T \\ \Delta_{10} \end{bmatrix} X_{1ij}$.

Define $\mathcal{E}(A) = E(A|u_{1j}^*, Y_j)$, $\mathcal{V}(A) = var(A|u_{1j}^*, Y_j)$ and $\mathcal{C}(A, B) = cov(A, B|u_{1j}^*, Y_j)$. We have multivariate normal $f(\nu_{0j}, \nu_{2j}, r_{ij}^* | u_{1j}^*, Y_j)$ for

$$\begin{aligned} \mathcal{E}(\nu_{0j}) &= T_{01} \tau_{11}^{-1} u_{1j}^* + \Delta_j^{-1} \left( \sum_i O_{ij}^T \Sigma_{ij}^{-1} e_{o1ij} + \Omega_{00j}^{-1} T_{02|1} O_{2j}^T T_{22|1j}^{-1} e_{o2j} \right), \\ \mathcal{E}(\nu_{2j}) &= T_{21} \tau_{11}^{-1} u_{1j}^* + T_{22|1} O_{2j}^T T_{22|1j}^{-1} e_{o2j} + \Omega_{20j} \Omega_{00j}^{-1} \left[ \mathcal{E}(\nu_{0j}) - T_{01} \tau_{11}^{-1} u_{1j}^* - T_{02|1} O_{2j}^T T_{22|1j}^{-1} e_{o2j} \right], \\ \mathcal{E}(r_{ij}^*) &= \Sigma_j^* A_{OOij} [d_{ij}^* - \mathcal{E}(\nu_{0j})], \qquad \mathcal{V}(r_{ij}^*) = \Sigma_j^* - \Sigma_j^* (A_{OOij} - A_{OOij} \Delta_j^{-1} A_{OOij}) \Sigma_j^*, \\ \mathcal{V}(\nu_{0j}) &= \Delta_j^{-1}, \qquad \mathcal{C}(\nu_{0j}, \nu_{2j}) = \Delta_j^{-1} \Omega_{00j}^{-1} \Omega_{02j}, \qquad \mathcal{C}(\nu_{0j}, r_{ij}^*) = -\Delta_j^{-1} A_{OOij} \Sigma_j^*, \\ \mathcal{V}(\nu_{2j}) &= \Omega_{22j} - \Omega_{20j} (\Omega_{00j}^{-1} - \Omega_{00j}^{-1} \Delta_j^{-1} \Omega_{00j}^{-1}) \Omega_{02j}, \qquad \mathcal{C}(\nu_{2j}, r_{ij}^*) = \Omega_{20j} \Omega_{00j}^{-1} \mathcal{C}(\nu_{0j}, r_{ij}^*). \end{aligned}$$

Let $\tilde{A}_{1ij}^* = A_{1ij}^* - \Delta_{002} X_{2j} - \nu_{1j}$, $\delta_{10} = vec(\Delta_{10}^T)$ and $\beta_j = (I_{p_1+1+p_2} \otimes X_{2j}^T) \gamma_\beta + u_j^*$ for

$\gamma_\beta = vec[\gamma_{00}\ \gamma_{10}\ \Delta_{00}^T\ \Delta_2^T]$ and $u_j^* = [u_{0j}^*\ u_{1j}^*\ \boldsymbol{\nu}_j^T]^T$. The expected CD MLEs are

$$\begin{aligned}
\hat{\gamma}_{20} &= \gamma_{20} + \left(\sum_j E\left[\sum_i \mathcal{E}(\tilde{A}_{1ij}^* \tilde{A}_{1ij}^{*T})|Y_j\right]\right)^{-1} \sum_j E\left[\sum_i \mathcal{E}(e_{ij}^* \tilde{A}_{1ij}^*)|Y_j\right], \\
\hat{\gamma}_{30} &= \gamma_{30} + \left(\sum_j \sum_i X_{1ij} X_{1ij}^T\right)^{-1} \sum_j \sum_i X_{1ij} E\left[\mathcal{E}(e_{ij}^*)|Y_j\right], \\
\hat{\delta}_{10} &= \delta_{10} + vec\left[\left(\sum_j \sum_i E\left[\mathcal{E}(\boldsymbol{\epsilon}_{ij}^*)|Y_j\right] X_{1ij}^T\right)\left(\sum_j \sum_i X_{1ij} X_{1ij}^T\right)^{-1}\right] \\
\hat{\gamma}_\beta &= \gamma_\beta + vec\left[\left(\sum_j E[\mathcal{E}(u_j^*)|Y_j] X_{2j}^T\right)\left(\sum_j X_{2j} X_{2j}^T\right)^{-1}\right] \\
\hat{\sigma}^2 &= \sum_j E\left[\sum_i \mathcal{E}(e_{ij}^{*2})|Y_j\right]/N, \ \ \hat{\Sigma}_\epsilon = \sum_j E\left[\sum_i \mathcal{E}(\boldsymbol{\epsilon}_{ij}^* \boldsymbol{\epsilon}_{ij}^{*T})|Y_j\right]/N, \\
\hat{\tau} &= \sum_j E[\mathcal{E}(u_j^* u_j^{*T})|Y_j]/J
\end{aligned}$$

given $\theta$ where $\mathcal{E}(e_{ij}^*) = B_{1j}^{*T}\mathcal{E}(r_{ij}^*)$ and $\mathcal{E}(e_{ij}^{*2}) = B_{1j}^{*T}\mathcal{E}(r_{ij}^* r_{ij}^{*T})B_{1j}^*$ for $B_{1j}^{*T} = [1\ -B_{1j}^T]$.

## Appendix C: Numerical Integration by AGHQ

Let $f(u_{1j}^*) = h(Y_j|u_{1j}^*)\phi(u_{1j}^*; 0, \tau_{11})$ be a function of $u_{1j}^*$. Given $\tilde{u}_{1j}^* = E(u_{1j}^*|Y_j)$, $V_{u1j} = var(u_{1j}^*|Y_j) = L_{u1j}^2/2$, $Q$-point weights $(w_1, \cdots, w_Q)$ and abscissas $(a_1, \cdots, a_Q)$,

$$h(Y_j) = \int \frac{\phi(u_{1j}^*; \tilde{u}_{1j}^*, V_{u1j})}{\phi(u_{1j}^*; \tilde{u}_{1j}^*, V_{u1j})} f(u_{1j}^*) du_{1j}^* \ \approx \ L_{u1j} \sum_{k=1}^{Q} w_k e^{a_k^2} f(z_{kj}), \tag{37}$$

for $z_{kj} = L_{u1j} a_k + \tilde{u}_{1j}^*$. The $g(u_{1j}^*|Y_j) = f(u_{1j}^*)/h(Y_j)$ is approximately $\phi(u_{1j}^*; \tilde{u}_{1j}^*, V_{u1j})$ for large cluster sizes $n_j$ by the Bayesian central limit theorem such that $f(u_{1j}^*) \propto \phi(u_{1j}^*; \tilde{u}_{1j}^*, V_{u1j})$ produces well approximated $h(Y_j)$ by a low degree polynomial. The approximation is exact if $f(u_{1j}^*)$ is a $2Q-1$ degree polynomial in $u_{1j}^*$ (Pinheiro and Bates 1995; Rabe-Hesketh et al.

2002; Carlin and Louis 2009). Likewise, for $E(S_{CDj}|u^*_{1j}, Y_j)$ closed-form,

$$E\left(S_{CDj}|Y_j\right) = \int E(S_{CDj}|u^*_{1j}, Y_j)g(u^*_{1j}|Y_j)du^*_{1j} \approx \frac{L_{u1j}}{h(Y_j)}\sum_{k=1}^{Q} E(S_{CDj}|z_{kj}, Y_j)w_k e^{a_k^2} f(z_{kj}) \quad (38)$$

Let $Y^*_j = (Y_j, Y_{misj})$, and $\phi_\tau$ and $\phi_\Sigma$ be vectors of distinct elements of $\tau$ and $\Sigma_\epsilon$, respectively. The loglikelihood $l = \sum_j l_j$ and score $S = \sum_j S_j$ have summands

$$l_j = \ln h(Y_j) = \ln \int g_j dY_{misj} d\nu_{0j} du^*_{1j}, \qquad S_j = \frac{\partial l_j}{\partial \theta} = E\left[\mathcal{E}\left(\frac{\partial \ln g_j}{\partial \theta}\right)|Y_j\right].$$

for $g_j = \prod_i f(R^*_{ij}|\mathbf{A}^*_{1ij}, u^*_j; \gamma_{20}, \gamma_{30}, \sigma^2) f(\mathbf{A}^*_{1ij}|u^*_j; \delta_{10}, \phi_\Sigma)\phi(u^*_j; \gamma_\beta, \tau)$ from Equations (35). Let $E = \frac{\partial vec\tau}{\partial \phi_\tau^T}$ and $F = \frac{\partial vec\Sigma_\epsilon}{\partial \phi_\Sigma^T}$. The $\mathcal{E}\left(\frac{\partial \ln g_j}{\partial \theta}\right)$ stacks

$$\mathcal{E}\left(\frac{\partial \ln g_j}{\partial \gamma_{20}}\right) = \sigma^{-2}\sum_i \mathcal{E}(e^*_{ij}\tilde{A}^*_{1ij}), \qquad \mathcal{E}\left(\frac{\partial \ln g_j}{\partial \gamma_{30}}\right) = \sigma^{-2}\sum_i \mathcal{E}(e^*_{ij})X_{1ij},$$

$$\mathcal{E}\left(\frac{\partial \ln g_j}{\partial \delta_{10}}\right) = vec\left(X_{1ij}\sum_i \mathcal{E}(\boldsymbol{\epsilon}^{*T}_{ij})\Sigma_\epsilon^{-1}\right), \qquad \mathcal{E}\left(\frac{\partial \ln g_j}{\partial \sigma^2}\right) = \frac{1}{2}A_{\sigma j}, \quad \mathcal{E}\left(\frac{\partial \ln g_j}{\partial \phi_\Sigma}\right) = \frac{1}{2}F^T A_{\Sigma j},$$

$$\mathcal{E}\left(\frac{\partial \ln g_j}{\partial \gamma_\beta}\right) = vec\left(X_{2j}\mathcal{E}(u^{*T}_j)\tau^{-1}\right), \qquad \mathcal{E}\left(\frac{\partial \ln g_j}{\partial \phi_\tau}\right) = \frac{1}{2}E^T vec[\tau^{-1}\mathcal{E}(u^*_j u^{*T}_j)\tau^{-1} - \tau^{-1}]$$

for $A_{\sigma j} = \sigma^{-4}\sum_i \mathcal{E}(e^{*2}_{ij}) - n_j\sigma^{-2}$ and $A_{\Sigma j} = vec\left(\Sigma_\epsilon^{-1}\sum_i \mathcal{E}(\boldsymbol{\epsilon}^*_{ij}\boldsymbol{\epsilon}^{*T}_{ij})\Sigma_\epsilon^{-1} - n_j\Sigma_\epsilon^{-1}\right)$. We compute $S_j$ also by AGHQ for $var(\hat{\theta}) \approx \left(\sum_j S_j S_j^T\right)^{-1}$ (Hedeker and Gibbons 1994; Raudenbush et al. 2000; Olsen and Schafer 2001). Section 7 shows good approximation for the sample sizes analyzed in this paper.

## Appendix D: Translating model (35) to $\hat{\theta}^*_{(19)}$

Define $\beta_{kj} = \gamma^T_{k0}X_{2j} + u^*_{kj}$, $\beta_{Cj} = \delta^T_{001}X_{2j} + \nu_j$, $cov(u^*_{kj}, u^*_{k'j}|X_{2j}) = \tau_{kk'}$, $cov(u^*_{kj}, \nu_j|X_{2j}) = \tau_{k\nu}$ and $var(\nu_j|X_{2j}) = \tau_{\nu\nu}$ for $k, k' = 0, 1$. With $X_{2j}$ marginalized out, $\beta_j = [\beta_{0j}\ \beta_{1j}\ \beta_{Cj}]^T \sim$

$$N\left(\begin{bmatrix}\gamma^T_{00}\\ \gamma^T_{10}\\ \delta^T_{121}\end{bmatrix}EX_2, \begin{bmatrix}t_{00} & t_{01} & t_{0\nu}\\ & t_{11} & t_{1\nu}\\ & & t_{\nu\nu}\end{bmatrix} = \begin{bmatrix}\tau_{00} & \tau_{01} & \tau_{0\nu}\\ & \tau_{11} & \tau_{1\nu}\\ & & \tau_{\nu\nu}\end{bmatrix} + \begin{bmatrix}\gamma^T_{00}\\ \gamma^T_{10}\\ \delta^T_{001}\end{bmatrix}VX_2[\gamma_{00}\ \gamma_{10}\ \delta_{001}]\right)$$

for $E(X_{2j}) = EX_2$ and $var(X_{2j}) = VX_2$. Let $\tilde{\beta}_{Cj} = \beta_{Cj} - \delta_{001}^T EX_2 \sim N(0, t_{\nu\nu})$ to find

$$\beta_{kj}|\tilde{\beta}_{Cj} \sim \left(\gamma_{k0}^* + \gamma_{k1}^* \tilde{\beta}_{Cj}, \tau_{kk}^*\right) \tag{39}$$

for $\gamma_{k0}^* = \gamma_{k0}^T EX_2$, $\gamma_{k1}^* = t_{k\nu}/t_{\nu\nu}$ and $cov(\beta_{kj}, \beta_{k'j}|\tilde{\beta}_{Cj}) = \tau_{kk'}^* = t_{kk'} - \gamma_{k1}^* t_{\nu\nu} \gamma_{k'11}^*$.

Within cluster $j$ given $u_j^*$, denote $\tilde{\boldsymbol{A}}_{1ij}^* = \boldsymbol{A}_{1ij}^* - \Delta_{00} X_{2j} - \boldsymbol{\nu}_{1j} = \Delta_{10} X_{1ij} + \boldsymbol{\epsilon}_{ij}$. Marginalizing $X_{1ij}$ out using $EX_1 = E(X_{1ij})$ and $VX_1 = var(X_{1ij})$, we find $f(Y_{ij}^*|u_j^*)$ in

$$\tilde{\boldsymbol{A}}_{1ij}^* = \Delta_{10} EX_1 + \tilde{\boldsymbol{\epsilon}}_{ij}^*, \qquad R_{ij} = \beta_{0j} + (B_{1j}^T \Delta_{10} + \gamma_{30}^T) EX_1 + B_{1j}^T \tilde{\boldsymbol{\epsilon}}_{ij}^* + \tilde{e}_{ij}^*$$

for $\begin{bmatrix} \tilde{e}_{ij}^* \\ \tilde{\boldsymbol{\epsilon}}_{ij}^* \end{bmatrix} \sim N\left(0, \begin{bmatrix} \sigma_{ee}^* & \Sigma_{e\epsilon}^* \\ \Sigma_{\epsilon e}^* & \Sigma_{\epsilon\epsilon}^* \end{bmatrix}\right)$ where $\sigma_{ee}^* = \gamma_{30}^T VX_1 \gamma_{30} + \sigma^2$, $\Sigma_{\epsilon e}^* = \begin{bmatrix} \sigma_{ce} \\ \Sigma_{1e} \end{bmatrix} = \begin{bmatrix} \delta_{101}^T \\ \Delta_{102} \end{bmatrix} VX_1 \gamma_{30}$ and $\Sigma_{\epsilon\epsilon}^* = \begin{bmatrix} \sigma_{cc}^* & \Sigma_{c1}^* \\ \Sigma_{1c}^* & \Sigma_{11}^* \end{bmatrix} = \begin{bmatrix} \delta_{101}^T VX_1 \delta_{101} + \sigma_{cc} & \delta_{101}^T VX_1 \Delta_{102}^T + \Sigma_{c1} \\ \Delta_{102} VX_1 \delta_{101} + \Sigma_{1c} & \Delta_{102} VX_i \Delta_{102}^T + \Sigma_{11} \end{bmatrix}$.
Consequently, $R_{ij}|\beta_j, \epsilon_{ij} \sim N[E(R_{ij}^*|\beta_j, \epsilon_{ij}), \sigma^{'2}]$ for

$$\begin{aligned} E(R_{ij}^*|\beta_j, \epsilon_{ij}) &= \beta_{0j} + (B_{1j}^T \Delta_{10} + \gamma_{30}^T) EX_1 + (\beta_{1j} + \gamma_{20}^T \Sigma_{1c}^*/\sigma_{cc}^* + \sigma_{ec}^*/\sigma_{cc}^*)\epsilon_{ij}, \\ \sigma^{'2} &= \gamma_{20}^T (\Sigma_{11}^* - \Sigma_{1c}^* \Sigma_{c1}^*/\sigma_{cc}^*)\gamma_{20} + 2\gamma_{20}^T (\Sigma_{1e}^* - \Sigma_{1c}^* \sigma_{ce}^*/\sigma_{cc}^*) + (\sigma_{ee}^* - \sigma_{ce}^{*2}/\sigma_{cc}^*). \end{aligned} \tag{40}$$

As explained by Section 4.1, Equations (39)-(40) result in a mixed model (19) for

$$\begin{aligned} \gamma_{00} = \gamma_{00}^* + (\gamma_{10}^* \delta_{101}^T + \gamma_{20}^T \Delta_{102} + \gamma_{30}^T) EX_1, \quad \gamma_{01} = \gamma_{01}^* + \gamma_{11}^* \delta_{101}^T EX_1, \\ \gamma_{10} = \gamma_{10}^* + \gamma_{20}^T \Sigma_{1c}^*/\sigma_{cc}^* + \sigma_{ec}^*/\sigma_{cc}^*, \qquad \gamma_{11} = \gamma_{11}^*, \qquad \sigma^2 = \sigma^{'2} \\ \tau_{00|\nu} = \tau_{00}^* + 2\tau_{01}^* \delta_{101}^T EX_1 + \tau_{11}^* (\delta_{101}^T EX_1)^2, \quad \tau_{01|\nu} = \tau_{01}^* + \tau_{11}^* \delta_{101}^T EX_1, \quad \tau_{11|\nu} = \tau_{11}^*. \end{aligned} \tag{41}$$

Estimating $(EX_1, VX_1, EX_2, VX_2)$ from sample, we find $\hat{\theta}_{(19)}^*$ above and compute $var(\hat{\theta}_{(19)}^*)$ by the delta method. Translations (41) simplify if $EX_1 = 0$ (e.g., $EX_1 = E(X_{1ij} - \bar{X}_{1j})$). In Section 6, we used the translations for $X_{2j} = 1$ and $X_{1ij} = 0$.

## Appendix E: Compatible Gibbs Sampler without a PKRE

Without the merit of a PKRE, a Bayesian joint distribution based on Equations (19) is

$$f(R^*_{ij}|C^*_{ij},\nu_j,u_{0j},u_{1j},\theta)f(u_{0j},u_{1j}|\theta)f(C^*_{ij}|\nu_j,\theta)\phi(\nu_j;0,\tau_{\nu\nu})p(\theta)$$

for a prior $p(\theta)$ and $\theta=\theta^*_{(19)}$. Our scientific model is an analytic integral

$$f_1(R^*_{ij}|C^*_{ij},\nu_j,\theta)=\int\int f(R^*_{ij}|C^*_{ij},\nu_j,u_{0j},u_{1j},\theta)f(u_{0j},u_{1j}|\theta)du_{0j}du_{1j}.$$

The Gibbs sampler of Enders et al. (2020) may be modified to be compatible with our scientific model conditional on a latent covariate $\nu_j$ by sampling i) $\nu_j$ from a compatible posterior

$$p(\nu_j|\cdot) \;=\; \frac{\prod_i f(R^*_{ij}|C^*_{ij},\nu_j,u_{0j},u_{1j},\theta)f(C^*_{ij}|\nu_j,\theta)\phi(\nu_j;0,\tau_{\nu\nu})}{\int\prod_i f(R^*_{ij}|C^*_{ij},\nu_j,u_{0j},u_{1j},\theta)f(C^*_{ij}|\nu_j,\theta)\phi(\nu_j;0,\tau_{\nu\nu})d\nu_j}$$

for the denominator approximated by the MCMC integration, and ii) a missing $C^*_{ij}$ from a compatible normal posterior $p(C^*_{ij}|\cdot)\propto f(R^*_{ij}|C^*_{ij},\nu_j,u_{0j},u_{1j},\theta)f(C^*_{ij}|\nu_j,\theta)$ with

$$E(C^*_{ij}|\cdot)=\delta+\nu_j+\frac{\beta_{1j}\sigma_{cc}}{\beta^2_{1j}\sigma_{cc}+\sigma^2}(R^*_{ij}-\beta_{0j}),\qquad var(C^*_{ij}|\cdot)=\frac{\sigma_{cc}\sigma^2}{\beta^2_{1j}\sigma_{cc}+\sigma^2}$$

by Equations (15) for $\beta_{0j}=\gamma_{00}+\gamma_{01}\nu_j+u_{0j}$ and $\beta_{1j}=\gamma_{10}+\gamma_{11}\nu_j+u_{1j}$.

# References


[1] Arnold, B. C. and Press, S. J. (1989). Compatible conditional distributions. *Journal of the American Statistical Association*, **84**, 152-156.

[2] Bartlett, J. W., Seaman, S. R., White, I. R., and Carpenter, J. R. (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research*, **24**, 462-487.

[3] Bates, D., Maechler, M., Bolker, B. and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, **67(1)**, 1-48. doi:10.18637/jss.v067.i01.

[4] Carlin, B. P. and Louis, T. A. (2009). *Bayesian methods for data analysis*. 3rd ed. Boca Raton, FL: CRC press.

[5] Collins, L.M., Schafer, J.L. and Kam, C. (2003). A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures. *Psychological Methods*, **6**, 330-351.

[6] Dempster, AP., Laird, NM., and Rubin, DB. (1977). Maximum Likelihood From Incomplete Data via the EM Algorithm. *JRSS, Series B*, **76**, 1-38.

[7] Dempster, AP., Rubin, DB. and Tsutakawa, RK. (1981). Estimation in Covariance Components Models. *JASA*, **76**, 341-353.

[8] Enders, C. K., Mistler, S. A., and Keller, B. T. (2016). Multilevel Multiple Imputation: A Review and Evaluation of Joint Modeling and Chained Equations Imputation. *Psychological Methods*, **21**, 222-240.

[9] Enders, C. K., Du, H., and Keller, B. T. (2020). A model-based imputation procedure for multilevel regression models with random coefficients, interaction effects, and nonlinear terms. *Psychological Methods*, **25(1)**, 88-112.

[10] Erler, N. S., Rizopoulos, D., Rosmalen, J., Jaddoe, V. W., Franco, O. H. and Lesaffre, E. M. (2016). Dealing with missing covariates in epidemiologic studies: A comparison between multiple imputation and a full Bayesian approach. *Statistics in Medicine*, **35**, 2955-74.

[11] Erler, N. S., Rizopoulos, D., Jaddoe, V. W., Franco, O. H. and Lesaffre, E. M. (2019). Bayesian imputation of time-varying covariates in linear mixed models. *Statistical Methods in Medical Research*, **28**, 555-568.

[12] Firebaugh, G. (1978). A rule for inferring individual-level relationships from aggregate data. *American Sociological Review*, **43**, 557–572.

[13] Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457-472.

[14] Goldstein, H., Carpenter, J., Kenward, M., and Levin, K. (2009). Multilevel models with multivariate mixed response types, *Statistical Modellng*, **9**, 173-197.

[15] Goldstein, H., Carpenter, J. R., and Browne, W. J. (2014). Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non-linear terms. it Journal of the Royal Statistical Society Series A-Statistics in Society, **177**, 553-564.

[16] Grund, S., Lüdtke, O. and Robitzsch, A. (2018). Multiple imputation of missing data for multilevel models: Simulations and recommendations. *Organizational Research Methods*, **21**, 111–149.

[17] Grund, S., Lüdtke, O. and Robitzsch, A. (2021). Multiple imputation of missing data in multilevel models with the R package mdmb: A flexible sequential modeling approach. *Behavior Research Methods*, **53**, 2631–49.

[18] Hedeker, D. and Gibbons, R.D. (1994). A Random-Effects Ordinal Regression Model for Multilevel Analysis, *Biometrics*, **50**, 933-944.

[19] Ibrahim, J. G., Chen, M. H. and Lipsitz, S. R. (2002). Bayesian methods for generalized linear models with covariates missing at random. *Canadian Journal of Statistics/Revue Canadienne De Statistique*, **30**, 55-78.

[20] Keller, B. T. and Enders, C. K. (2021). *Blimp user's guide* (Version 3). www.appliedmissingdata.com/multilevel-imputation.html

[21] Kim, S., Sugar, C. A., and Belin, T. R. (2015). Evaluating model-based imputation methods for missing covariates in regression models with interactions. *Statistics in Medicine*, **34**, 1876-1888.

[22] Lee, V.E. and Bryk, A.S. (1989). A Multilevel Model of the Social Distribution of High School Achievement. *Sociology of Education*, **62**, 172-192.

[23] Lindley, D. V. and Smith, A. F. M. (1972). Bayes Estimates for the Linear Model, *JRSS, Series B*, **34**, 1-41.

[24] Little, RJA. and Rubin, DB. (2002). *Statistical Analysis with Missing Data*, New York: Wiley.

[25] Liu, C., Rubin, B.D. and Wu, Y. (1998). Parameter Expansion to Accelerate EM: The PX-EM Algorithm. *Biometrika*, **85**, 755-770.

[26] Liu, M., Taylor, JMG. and Belin, TR. (2000). Multiple Imputation and Posterior Simulation for Multivariate Missing Data in Longitudinal Studies. *Biometrics*, **56** 1157-63.

[27] Liu, J., Gelman, A., Hill, J., Su, Y., and Kropko, J. (2014). On the stationary distribution of iterative imputations. *Biometrika*, **101**, 155-173.

[28] Miyazaki, Y. and Frank, K. A. (2006). A hierarchical linear model with factor analysis structure at level 2. *Journal of Educational and Behavioral Statistics*, **31**, 125-156.

[29] Naylor, JC. and Smith, AFM. (1982). Applications of a Method for the Efficient Computation of Posterior Distributions, *Applied Statistics*, **31(3)**, 214-225.

[30] Nomi, T. and Raudenbush, S. W. (2016). Making a Success of "Algebra for All:" the Impact of Extended Instructional Time and Classroom Peer Skill in Chicago. *Educational Evaluation and Policy Analysis*, **38(2)**, 431-451.

[31] Olsen, MK. and Schafer, JL. (2001). A Two-Part Random-Effects Model for Semicontinuous Longitudinal Data, *JASA*, **96**, 730-745.

[32] Piketty, T. (2014). *Capital in the 21st Century*, Harvard University Press Cambridge, MA.

[33] Pinheiro, JC. and Bates, DM. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model, *JCGS*, **4**, 12-35.

[34] R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[35] Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, **2(1)**, 1-21.

[36] Raghunathan, T., Lepkowski, J., Van Hoewyk, J. and Solenberger, P. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology*, **27**, 85-95.

[37] Raudenbush, S.W. and Bryk A.S. (1986). Hierarchical Model for Studying School Effects. *Sociology of Education*, 1-17.

[38] Raudenbush, SW., Yang, M., Yosef, M. (2000). Maximum Likelihood for Generalized Linear Models with Nested Random Effects via High-Order, Multivariate Laplace Approximation. *JCGS*, **9(1)**, 141-157.

[39] Raudenbush, SW. and Bryk, AS. (2002). *Hierarchical Linear Models*. Newbury Park, CA: Sage.

[40] Ren, C. and Shin, Y. (2016) Longitudinal Latent Variable Models Given Incompletely Observed Biomarkers and Covariates, *Statistics in Medicine*, **35**, 4729-45.

[41] Rivera-Batiz, F. L. (1982). International migration, non-traded goods and economic welfare in the source country. *Journal of Development Economics*, **11(1)**, 81-90.

[42] Rockwood, N. J. (2020). Maximum likelihood estimation of multilevel structural equation models with random slopes for latent covariates. *Psychometrika*, **85**, 275–300

[43] Rubin, D.B. (1976). Inference and missing data. *Biometrika*, **63**, 581-592.

[44] Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. J. Wiley & Sons, New York.

[45] Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

[46] Schafer, J.L. and Yucel, RM. (2002). Computational Strategies for Multivariate Linear Mixed-Effects Models with Missing Values. *Journal of Computational and Graphical Statistics*, **11**, 437-457.

[47] Shin, Y. and Raudenbush, SW. (2007). Just-Identified Versus Over-Identified Two-Level Hierarchical Linear Models with Missing Data. *Biometrics*, **63**, 1262-68.

[48] Shin, Y. and Raudenbush, SW. (2010). A Latent Cluster Mean Approach to The Contextual Effects Model with Missing Data. *Journal of Educational and Behavioral Statistics*, **35**, 26-53.

[49] Shin, Y. and Raudenbush, SW. (2013) Efficient Analysis of Q-Level Nested Hierarchical General Linear Models Given Ignorable Missing Data. *The International Journal of Biostatistics*, **9(1)**, 109-133.

[50] Stram, D.O. and Lee, J. (1994). Variance Components Testing in the Longitudinal Mixed Effects Model. *Biometrics*, **50**, 1171-77.

[51] Sun, X., Shin, Y., Lafata, J.E. and Raudenbush, S.W. (2023), Variability in Causal Effects and Noncompliance in a Multisite Trial: Estimation of a Bivariate Hierarchical Generalized Linear Model, Submitted.

[52] Tourangeau, K., Nord, C., Lê, T., Sorongon, AG., and Najarian, M. (2009). *Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), Combined User's Manual for the ECLS-K Eighth-Grade and K-8 Full Sample Data Files and Electronic*

*Codebooks* (NCES 2009-004). National Center for Education Statistics, Institute of Education Sciences, US Department of Education. Washington, D.C.

[53] van Buuren, S., Brand, J., Groothuis-Oudshoorn, C., and Rubin, D. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, **76**, 1049-64.

[54] Willms, J. D. (1986). Social class segregation and its relationship to pupils' examination results in Scotland. *American Sociological Review*, **51**, 224–241.