Efficient Analysis of Q-Level Nested Hierarchical General Linear Models Given Ignorable Missing Data

Yongyun Shin and Stephen W. Raudenbush

The International Journal of Biostatistics, v9 n1 p109-133 2013

Abstract

This paper extends single-level missing data methods to efficient estimation of a Q-level nested hierarchical general linear model given ignorable missing data with a general missing pattern at any of the Q levels. The key idea is to reexpress a desired hierarchical model as the joint distribution of all variables including the outcome that are subject to missingness, conditional on all of the covariates that are completely observed; and to estimate the joint model under normal theory. The unconstrained joint model, however, identifies extraneous parameters that are not of interest in subsequent analysis of the hierarchical model, and that rapidly multiply as the number of levels, the number of variables subject to missingness, and the number of random coefficients grow. Therefore, the joint model may be extremely high dimensional and difficult to estimate well unless constraints are imposed to avoid the proliferation of extraneous covariance components at each level. Furthermore, the over-identified hierarchical model may produce considerably biased inferences. The challenge is to represent the constraints within the framework of the Q-level model in a way that is uniform without regard to Q; in a way that facilitates efficient computation for any number of Q levels; and also in a way that produces unbiased and efficient analysis of the hierarchical model. Our approach yields Q-step recursive estimation and imputation procedures whose qth step computation involves only level-q data given higher-level computation components. We illustrate the approach with a study of the growth in body mass index analyzing a national sample of elementary school children.

KEY WORDS: Child Health; Hierarchical General Linear Model; Ignorable Missing Data; Maximum Likelihood; Multiple Imputation.

Acknowledgements: This research was supported by the Institute of Education Sciences, U.S. Department of Education, through Grants R305D090022 to NORC and R305D130033 to VCU. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. The first author was also supported by grant 1U01HL101064 from the National Institutes of Health.

1 Introduction

A seminal contribution to statistical methodology is the development of efficient methods for handling missing data within the framework of a general linear model (GLM, Rubin, 1976, 1987, Dempster, Laird, and Rubin, 1977, Meng, 1994, Schafer, 1997, 2003, Little and Rubin, 2002). These methods provide efficient estimation of the GLM given incomplete data. In particular, model-based multiple imputation now provides state-of-the-art methods for handling missing data (Rubin, 1987). These approaches are founded on a comparatively mild assumption in many applications that missing data are ignorable (Rubin, 1976, Little and Rubin, 2002).

This paper extends the methodology to an arbitrary Q-level hierarchical GLM where lower-level units are nested within higher-level units (Raudenbush and Bryk, 2002, Goldstein, 2003). Many multilevel observational studies and controlled experiments produce missing data. In cluster-randomized experiments, the dominant design involves the random assignment of whole schools, hospitals, or communities, rather than students, patients, or adults to treatments (Bingenheimer and Raudenbush, 2004). Multilevel analysis is pervasive in health, education, and social science studies (Datar and Sturm, 2004, Gable, Chang, and Krull, 2007, Danner, 2008, Shin and Raudenbush, 2010). Surveys involve multi-stage sampling designs (Tourangeau, Nord, Lê, Sorongon, and Najarian, 2009). A ubiquitous problem is that explanatory as well as outcome variables may be subject to missingness at any of the levels.

In longitudinal studies, hierarchical data subject to missingness may be estimated by maximum likelihood (ML) in a structural equation model (SEM) where latent means include missing data (Allison, 1987, Muthén, Kaplan, and Hollis, 1987, Muthén, 1993, Arbuckle, 1996, Enders and Peugh, 2004). SEM software such as Mplus (Muthén and Muthén, 2010), Amos (Arbuckle, 2003) and EQS (Bentler, 2007) performs ML estimation of such models. When these models are formulated by multi-group analysis, the number of groups is the number of missing patterns (Allison, 1987, Muthén et al., 1987, Muthén, 1993).

Recent advances have extended the single-level methods to multilevel ignorable missing data. Liu, Taylor, and Belin (2000) considered Bayes inference to longitudinal designs having a fixed within-subject design with repeated measures at level 1 nested within persons at level 2 where the data are missing at both levels. Schafer and Yucel (2002) developed Bayes and ML inference for a broader class of two-level designs in which the level-1 design could vary across level-2 units with level-1 data subject to missingness. Goldstein and Browne (2002, 2005) took a Bayesian approach to a two-level factor analysis where missing outcomes were imputed by a Gibbs sampling step. Shin and Raudenbush (2007, 2010) extended these methods to a two-level model where the outcome and covariates may have missing values at both levels. Shin and Raudenbush (2011) and Shin (2012) illustrated an efficient ML method to estimate a three-level model with incomplete data. To estimate a three-level hierarchical linear model with incomplete data, Yucel (2008) modified a single-level imputation method (Schafer, 1997, 1999) and a two-level imputation method (Schafer and Yucel, 2002) to carry out the Gibbs Sampler to sequentially impute cluster-level missing values, intermediate-level missing values given the multiply imputed cluster-level data and then, the lowest-level missing values given the multiply imputed data at higher levels. These advances guide us with continuous outcomes. Goldstein, Carpenter, Kenward, and Levin (2009) and Goldstein and Kounali (2009) used a Markov Chain Monte Carlo method to impute a mixture of continuous and discrete outcomes subject to missingness in a two-level model.

Shin and Raudenbush (2007) illustrated two ways to handle two-level missing data: direct ML estimation ($MLE\ on\ Y_{obs}$) and a two-stage procedure of multiple imputation followed by the second stage analysis of the multiply imputed data ($MLE\ on\ Y^{mi}$). This paper generalizes the two methods to an arbitrary number of Q levels and an arbitrary number of outcomes defined at any level. A key emphasis in this paper is the difference in logic and assumptions between the two methods. Using $MLE\ on\ Y_{obs}$, one first writes down a desired hierarchical model, then reparameterizes the model in terms of the joint distribution of outcome and covariates subject to missingness given the completely observed covariates. Great care must be taken so that the transformation is one-to-one in order to insure unbiased estimation (Shin and Raudenbush, 2007, 2010). This task generally requires the imposition of constraints on regression surfaces at each level to avoid the proliferation of covariance components at higher levels of the joint distribution. One challenge for this paper is to formulate these constraints within the framework of the Q-level model. We show that the unconstrained joint model identifies contextual effects (Raudenbush and Bryk, 2002) and interaction effects that are typically extraneous for the desired model. In contrast, $MLE\ on\ Y^{mi}$ generally implies that the imputation model should be unconstrained, allowing the data analyst to impose the desired constraints at the second stage when using conventional software to analyze the imputed data. A challenge with $MLE\ on\ Y^{mi}$ is that the need to avoid constraints at stage one may lead to the formulation of extremely high-dimensional imputation models that may be difficult to estimate well. The two methods have characteristic advantages and disadvantages. $MLE\ on\ Y_{obs}$ imposes more assumptions than does $MLE\ on\ Y^{mi}$, because $MLE\ on\ Y_{obs}$ imposes distributional assumptions on all variables subject to missingness while for $MLE\ on\ Y^{mi}$, such assumptions do not affect the observed data. Given the pluses and minuses, this paper considers a hybrid approach that combines the two methods.

Our aim is to formulate a Q-level model that unifies single- and multi-level models into a single expression, facilitating extension of existing missing data methods to an arbitrary number of levels of a linear model with efficient estimation and computation. The model has two representations: a hierarchical linear model for a response variable conditional on covariates, and a marginal model that represents the joint distribution of all variables - the response and covariates - that are subject to missingness conditional on the completely observed covariates. It is essential to clarify the relationship between these two models; in particular, the conditional model is always equivalent to the joint model after imposing certain constraints on the more general joint model. To clarify the needed constraints in a general Q-level setting, we find it revealing to re-parameterize both forms of the model such that all variables subject to missingness are decomposed orthogonally by level. In this model, random components are correlated within but uncorrelated between levels. The required constraints then fall out naturally for any number of levels.

The next section defines the joint model and decomposes all variables subject to missingness into orthogonal random components. Section 3 considers the problem of estimation and multiple imputation. The orthogonal decomposition is helpful. The key insight is that, if we stack the joint model by level, we can write down estimation formulas at level q using level-q data only given higher-level computation components that are uniform for all q even if we ignore all other-level data. This recursive representation yields efficient computations using conventional ML methods such as the EM algorithm (Dempster, Laird, and Rubin, 1977). Thus, orthogonal decomposition by level transforms a seemingly intractable computational problem into a sequence of familiar, solvable problems as described in Section 3. Section 4 introduces the conditional hierarchical linear model. We show that the joint model represents more parameters than desired in the hierarchical model, and describe how to constrain the joint model for identification of the desired hierarchical model. Incorporating the constraints is essential to avoid bias for $MLE\ on\ Y_{obs}$. For $MLE\ on\ Y^{mi}$, the constraints do not reduce bias but may nonetheless be practically necessary for computation involving high dimensional models. Analyzing data from a large, nationally representative longitudinal sample of children, we illustrate these methods to study predictors of the growth of body mass index (BMI) between ages 5 to 15 in Section 5. This is a three-level problem, with up to seven repeated measures on children who are sampled within their elementary schools. Section 6 concludes with a discussion of limitations and next steps in the Q-level research agenda.

2 Joint Model

All of the models described in this article are subsets of a multilevel p-variate model

$$Y = \mu + Zb \sim N(\mu, V), \qquad b \sim N(0, \Pi) \tag{1}$$

where every element of Y is subject to missingness, μ may be a linear function of completely observed covariates, Z is a matrix of completely observed covariates having random effects b and $V = Z\Pi Z^T$. For simplicity of exposition, we shall assume $\mu = 0$ in this section. We partition $Y = [Y_1^T\ Y_2^T \cdots Y_Q^T]^T$ such that Y_q is a vector of p_q variables at level q in a hierarchy of Q levels for $p = \sum_{q=1}^{Q} p_q$. In the case of $Q = 2$ where occasions are nested within children, for example, elements of Y_1 such as body mass index and daily TV viewing hours vary across occasions at the lower level 1; and elements of Y_2 such as years of highest parent education and birth weight vary among children at the higher level 2.

The variance covariance matrix V may be structured by the fact that Y_q varies at level q or higher. In the case of $Q = 2$, for example, the body mass index in Y_1 varies within as well as between children while the birth weight in Y_2 varies between children but not within a child. Thus, we partition $Z = \bigoplus_{q=1}^{Q} Z_q = diag\{Z_1, Z_2, \cdots, Z_Q\}$, a diagonal matrix having diagonal submatrices $(Z_1, Z_2, \cdots, Z_Q)$, and $b = [b_1^T\ b_2^T \cdots b_Q^T]^T$, and decompose Y_q orthogonally by level as

$$Y_q = Z_q b_q \sim N\left(0, V_{qq}\right), \qquad b_q \sim N\left(0, \Pi_{qq}\right) \tag{2}$$

for $Z_q = [Z_{qq}\ Z_{(q+1)q} \cdots Z_{Qq}]$, $b_q = [\epsilon_{qq}^T\ \epsilon_{(q+1)q}^T \cdots \epsilon_{Qq}^T]^T$ and $V_{qq} = Z_q \Pi_{qq} Z_q^T$ where Z_{rq} is a matrix of known covariates having level-r unit-specific random effects $\epsilon_{rq} \sim N(0, \pi_{qq}^r)$ for subscript and superscript r denoting the level of variation. The orthogonal random effects ϵ_{rq} are independent between levels so that $\Pi_{qq} = \bigoplus_{r=q}^{Q} \pi_{qq}^r$, but correlated within levels by $cov(\epsilon_{rq}, \epsilon_{rs}) = \pi_{qs}^r$ so that $cov(b_q, b_s) = \Pi_{qs} = \begin{bmatrix} 0 \\ \bigoplus_{r=s}^{Q} \pi_{qs}^r \end{bmatrix}$ for $q < s$. Then, $\Pi = [\Pi_{ij}]$, a matrix with (i, j) submatrices Π_{ij} for $i, j = 1, 2, \cdots, Q$. We may now express Equation

(2) as $Y_q = \sum_{r=q}^{Q} Z_{rq}\epsilon_{rq}$ having $cov(Y_q, Y_s) = V_{qs} = Z_q \Pi_{qs} Z_s^T = \sum_{r=s}^{Q} v_{qs}^r$ for $q \leq s$ where $v_{qs}^r = Z_{rq}\pi_{qs}^r Z_{rs}^T$, and decompose $V = [V_{ij}]$ for $i, j = 1, 2, \cdots, Q$ by the level of variation as

$$var \left(\begin{array}{c} Y_1 \\ Y_2 \\ \vdots \\ Y_Q \end{array} \right) = \left[\begin{array}{cccc} v_{11}^1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{array} \right] + \left[\begin{array}{cccc} v_{11}^2 & v_{12}^2 & \cdots & 0 \\ v_{21}^2 & v_{22}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{array} \right] + \cdots + \left[\begin{array}{cccc} v_{11}^Q & v_{12}^Q & \cdots & v_{1Q}^Q \\ v_{21}^Q & v_{22}^Q & \cdots & v_{2Q}^Q \\ \vdots & \vdots & \ddots & \vdots \\ v_{Q1}^Q & v_{Q2}^Q & \cdots & v_{QQ}^Q \end{array} \right]. \quad (3)$$

To reveal the orthogonal decomposition explicitly, we show all random components of the joint model (1) in Table 1. Row label Y_q indicates a vector of variables that are decomposed into the random components $b_q = [\epsilon_{qq}^T \, \epsilon_{(q+1)q}^T \cdots \epsilon_{Qq}^T]^T$ listed in the row. The random components $(b_1, b_2, \cdots, b_Q)$ enable us to write down estimation formulas at level q that are uniform for all q and that use level-q data only given higher-level computation components. This representation facilitates construction of efficient computation formulas as will be explained in Section 3. The column label shows the level at which the random components in the column vary. Table 1 lists random effects that are correlated within, but uncorrelated across columns or levels. Column q in Table 1 lists level-q unit-specific random effects $\epsilon_q \sim N(0, \pi_q)$ for

Table 1: All random components of Y in Equation (1).

	1	2	3	$\cdots$	Q
Y_1	ϵ_{11}	ϵ_{21}	ϵ_{31}	$\cdots$	ϵ_{Q1}
Y_2		ϵ_{22}	ϵ_{32}	$\cdots$	ϵ_{Q2}
Y_3			ϵ_{33}	$\cdots$	ϵ_{Q3}
$\vdots$				$\ddots$	$\vdots$
Y_Q					ϵ_{QQ}

$$\epsilon_q = \left[\begin{array}{c} \epsilon_{q1} \\ \epsilon_{q2} \\ \vdots \\ \epsilon_{qq} \end{array} \right], \quad \pi_q = \left[\begin{array}{cccc} \pi_{11}^q & \pi_{12}^q & \cdots & \pi_{1q}^q \\ \pi_{21}^q & \pi_{22}^q & \cdots & \pi_{2q}^q \\ \vdots & \vdots & \ddots & \vdots \\ \pi_{q1}^q & \pi_{q2}^q & \cdots & \pi_{qq}^q \end{array} \right]. \quad (4)$$

Thus, all random effects b of the joint model (1) may also be expressed as $(\epsilon_1, \epsilon_2, \cdots, \epsilon_Q)$, displayed vertically in Table 1. Their parameters are $\pi = (\pi_1, \pi_2, \cdots, \pi_Q)$. This expression is useful for deriving estimators as will be described in the next Section.

3 Estimation of the Joint Model

The orthogonal decomposition by level shown in Table 1 enables us to write down the joint model (1) as a familiar mixed linear model (2) for any level q. Next, we shall exploit the fact that if we stack these level-specific models such that there is an equation at level q and a second equation stacked at all levels higher than q, we can write down familiar estimation formulas that use level-q data only given higher-level computation components even when we completely ignore all data at other levels. Moreover, the estimation formulas remain uniform for all q to produce efficient computation formulas as will be explained below. Therefore, the orthogonal decomposition of the joint model (1) enables us to obtain general, recursive, and familiar estimation formulas for the Q-level problem.

3.1 Structure of Joint Model for All Data at or above Level q

Based on the model (2) at level q, we stack $Y^q = [Y_q^T \ Y_{q+1}^T \cdots Y_Q^T]^T$, $Z^q = \bigoplus_{r=q}^Q Z_r$ and $b^q = [b_q^T \ b_{q+1}^T \cdots b_Q^T]^T$ to express the joint model (1) at level q or above as

$$Y^q = Z^q b^q \sim N(0, V^{qq}), \qquad b^q \sim N(0, \Pi^{qq}) \tag{5}$$

for $Y^q = \begin{bmatrix} Y_q \\ Y^{q+1} \end{bmatrix}$, $Z^q = \begin{bmatrix} Z_q & 0 \\ 0 & Z^{q+1} \end{bmatrix}$, $b^q = \begin{bmatrix} b_q \\ b^{q+1} \end{bmatrix}$, $\Pi^{qq} = \begin{bmatrix} \Pi_{qq} & \Pi^{q(q+1)} \\ \Pi^{(q+1)q} & \Pi^{(q+1)(q+1)} \end{bmatrix}$ and $V^{qq} = Z^q \Pi^{qq} Z^{qT} = \begin{bmatrix} V_{qq} & V^{q(q+1)} \\ V^{(q+1)q} & V^{(q+1)(q+1)} \end{bmatrix}$ where $\Pi^{q(q+1)} = \begin{bmatrix} \Pi_{q(q+1)} \ \Pi_{q(q+2)} & \cdots & \Pi_{qQ} \end{bmatrix}$ and $V^{q(q+1)} = Z_q \Pi^{q(q+1)} Z^{(q+1)T} = \begin{bmatrix} V_{q(q+1)} \ V_{q(q+2)} & \cdots & V_{qQ} \end{bmatrix}$ for the transpose $Z^{(q+1)T}$ of Z^{q+1}. Note that we express Y^q, Π^{qq} and V^{qq} uniformly for all q to contain Y^{q+1}, $\Pi^{(q+1)(q+1)}$ and $V^{(q+1)(q+1)}$, respectively. For $Q = 3$, for example, $Y^3 = Z^3 b^3 \sim N(0, V^{33})$ for $Y^3 = Y_3$, $Z^3 = Z_3 = Z_{33}$, $b^3 = b_3 = \epsilon_{33}$, $\Pi^{33} = \Pi_{33} = \pi_{33}^3$, $V^{33} = V_{33} = Z_3 \Pi_{33} Z_3^T$; $Y^2 = Z^2 b^2 \sim N(0, V^{22})$ for $Y^2 = \begin{bmatrix} Y_2 \\ Y^3 \end{bmatrix}$, $Z^2 = \begin{bmatrix} Z_2 & 0 \\ 0 & Z^3 \end{bmatrix}$, $b^2 = \begin{bmatrix} b_2 \\ b^3 \end{bmatrix}$, $\Pi^{22} = \begin{bmatrix} \Pi_{22} & \Pi^{23} \\ \Pi^{32} & \Pi^{33} \end{bmatrix}$ and $V^{22} = \begin{bmatrix} V_{22} & V^{23} \\ V^{32} & V^{33} \end{bmatrix}$ where $Z_2 = [Z_{22} \ Z_{32}]$, $b_2 = \begin{bmatrix} \epsilon_{22} \\ \epsilon_{32} \end{bmatrix}$, $\Pi_{22} = \bigoplus_{r=2}^3 \pi_{22}^r$, $\Pi^{23} = \Pi_{23} = \begin{bmatrix} 0 \\ \pi_{23}^3 \end{bmatrix}$ and $V^{23} = V_{23} = Z_2 \Pi_{23} Z_3^T$; and $Y^1 = Z^1 b^1 \sim N(0, V^{11})$ for $Y^1 = \begin{bmatrix} Y_1 \\ Y^2 \end{bmatrix}$, $Z^1 = \begin{bmatrix} Z_1 & 0 \\ 0 & Z^2 \end{bmatrix}$, $b^1 = \begin{bmatrix} b_1 \\ b^2 \end{bmatrix}$, $\Pi^{11} = \begin{bmatrix} \Pi_{11} & \Pi^{12} \\ \Pi^{21} & \Pi^{22} \end{bmatrix}$ and $V^{11} = \begin{bmatrix} V_{11} & V^{12} \\ V^{21} & V^{22} \end{bmatrix}$ where $Z_1 = [Z_{11} \ Z_{21} \ Z_{31}]$, $b_1 = \begin{bmatrix} \epsilon_{11} \\ \epsilon_{21} \\ \epsilon_{31} \end{bmatrix}$, $\Pi_{11} = \bigoplus_{r=1}^3 \pi_{11}^r$, $\Pi^{12} = [\Pi_{12} \ \Pi_{13}]$ and $V^{12} = [V_{12} \ V_{13}]$ for $\Pi_{12} = \begin{bmatrix} 0 & 0 \\ \pi_{12}^2 & 0 \\ 0 & \pi_{12}^3 \end{bmatrix}$, $\Pi_{13} = \begin{bmatrix} 0 \\ 0 \\ \pi_{13}^3 \end{bmatrix}$, $V_{12} = Z_1 \Pi_{12} Z_2^T$ and $V_{13} = Z_1 \Pi_{13} Z_3^T$.

Equation (5) for $q = 1$ expresses the joint model (1) where Z^1 may include covariates having random effects at all levels. Often, the model (5) itself is of interest (Shin and Raudenbush, 2011, Shin, 2012). For a positive integer n, let I_n denote an n by n identity matrix. In this paper, we focus on estimation of the model (5) where $Z_{qq} = I_{p_q}$ for all q as many applications do. In what follows, we use the Kronecker product $A \otimes B$ that multiplies matrix B to each scalar element of matrix A (Magnus and Neudecker, 1988). In particular, $I_n \otimes B = diag\{B, \cdots, B\}$.

3.2 Estimation

In deriving estimators, it is essential to aggregate the stacked-up joint model (5) at the highest level-Q cluster m. We refer to subscript m as a unit at the highest level Q for $m = 1, 2, \cdots, N_Q$; N_{qm} as the number of units at level q nested within the cluster m; and $N_q = \sum_{m=1}^{N_Q} N_{qm}$, hereafter. Let $Y_{qm} = [Y_{q1m}^T \; Y_{q2m}^T \cdots Y_{qN_{qm}m}^T]^T$ and $\epsilon_{qsm} = [\epsilon_{qs1m}^T \; \epsilon_{qs2m}^T \cdots \epsilon_{qsN_{qm}m}^T]^T$ aggregate all Y_q and ϵ_{qs} in cluster m for $s \leq q$. The level-q unit-specific random effects in column q of Table 1 are aggregated to form the aggregated Equations (4) for cluster m

$$\epsilon_{qm} = \begin{bmatrix} \epsilon_{q1m} \\ \epsilon_{q2m} \\ \vdots \\ \epsilon_{qqm} \end{bmatrix}, \qquad var(\epsilon_{qm}) = \begin{bmatrix} I_{N_{qm}} \otimes \pi_{11}^q & I_{N_{qm}} \otimes \pi_{12}^q & \cdots & I_{N_{qm}} \otimes \pi_{1q}^q \\ I_{N_{qm}} \otimes \pi_{21}^q & I_{N_{qm}} \otimes \pi_{22}^q & \cdots & I_{N_{qm}} \otimes \pi_{2q}^q \\ \vdots & \vdots & \ddots & \vdots \\ I_{N_{qm}} \otimes \pi_{q1}^q & I_{N_{qm}} \otimes \pi_{q2}^q & \cdots & I_{N_{qm}} \otimes \pi_{qq}^q \end{bmatrix}.$$

Table 1 may also be aggregated to reveal all random components for cluster m. The aggregated table is of the same form as Table 1 with row label Y_{qm} replacing Y_q, the same column label and the vector ϵ_{qm} instead of ϵ_q in column q. The random components in row Y_{qm} of the table are $b_{qm} = [\epsilon_{qqm}^T \; \epsilon_{(q+1)qm}^T \cdots \epsilon_{Qqm}^T]^T$ to form the aggregated model (2)

$$Y_{qm} = Z_{qm} b_{qm} \sim N(0, V_{qqm}), \qquad b_{qm} \sim N(0, \Pi_{qqm}) \tag{6}$$

for a conformable matrix of known covariates $Z_{qm} = [Z_{qqm} \;\; Z_{(q+1)qm} \cdots Z_{Qqm}]$, $\Pi_{qqm} = \bigoplus_{r=q}^{Q} I_{N_{rm}} \otimes \pi_{qq}^r$ and $V_{qqm} = Z_{qm} \Pi_{qqm} Z_{qm}^T$ where $Z_{qqm} = I_{N_{qm} \times p_q}$. Then, $cov(b_{qm}, b_{sm}) = \Pi_{qsm} = \begin{bmatrix} 0 \\ \bigoplus_{r=s}^{Q} I_{N_{rm}} \otimes \pi_{qs}^r \end{bmatrix}$ for $q < s$, and $Y_{qm} = \sum_{r=q}^{Q} Z_{rqm} \epsilon_{rqm}$ has $cov(Y_{qm}, Y_{sm}) = V_{qsm} = Z_{qm} \Pi_{qsm} Z_{sm}^T = \sum_{r=s}^{Q} Z_{rqm} (I_{N_{rm}} \otimes \pi_{qs}^r) Z_{rsm}^T$ for $q \leq s$.

Next, we stack $Y_m^q = [Y_{qm}^T Y_{(q+1)m}^T \cdots Y_{Qm}^T]^T$, $Z_m^q = \bigoplus_{r=q}^{Q} Z_{rm}$ and $b_m^q = [b_{qm}^T \; b_{(q+1)m}^T \cdots b_{Qm}^T]^T$ to express the aggregated model (5) uniformly for all q as

$$Y_m^q = Z_m^q b_m^q \sim N(0, V_m^{qq}), \qquad b_m^q \sim N\left(0, \Pi_m^{qq}\right) \tag{7}$$

for $Y_m^q = \begin{bmatrix} Y_{qm} \\ Y_m^{q+1} \end{bmatrix}$, $Z_m^q = \begin{bmatrix} Z_{qm} & 0 \\ 0 & Z_m^{q+1} \end{bmatrix}$, $b_m^q = \begin{bmatrix} b_{qm} \\ b_m^{q+1} \end{bmatrix}$, $\Pi_m^{qq} = \begin{bmatrix} \Pi_{qqm} & \Pi_m^{q(q+1)} \\ \Pi_m^{(q+1)q} & \Pi_m^{(q+1)(q+1)} \end{bmatrix}$ and $V_m^{qq} = Z_m^q \Pi_m^{qq} Z_m^{qT} = \begin{bmatrix} V_{qqm} & V_m^{q(q+1)} \\ V_m^{(q+1)q} & V_m^{(q+1)(q+1)} \end{bmatrix}$ where $\Pi_m^{q(q+1)} = \left[\Pi_{q(q+1)m} \; \Pi_{q(q+2)m} \; \cdots \; \Pi_{qQm}\right]$ and $V_m^{q(q+1)} = Z_{qm} \Pi_m^{q(q+1)} Z_m^{(q+1)T} = \left[V_{q(q+1)m} \; V_{q(q+2)m} \; \cdots \; V_{qQm}\right]$. By expressing $Z_{qm} = [I_{N_{qm} \times p_q} \;\; Z_{-qqm}]$ having $b_{qm} = [\epsilon_{qqm}^T \;\; \epsilon_{-qqm}^T]^T$ for $Z_{-qqm} = [Z_{(q+1)qm} \cdots Z_{Qqm}]$ and $\epsilon_{-qqm} = [\epsilon_{(q+1)qm}^T \cdots \epsilon_{Qqm}^T]^T$ so that $\epsilon_{qqm} \sim N(0, I_{N_{qm}} \otimes \pi_{qq}^q)$, $\epsilon_{-qqm} \sim N(0, \Phi_{qqm})$ and $\Pi_m^{q(q+1)} = \begin{bmatrix} cov(\epsilon_{qqm}, b_m^{q+1}) \\ cov(\epsilon_{-qqm}, b_m^{q+1}) \end{bmatrix} = \begin{bmatrix} 0 \\ \Phi_m^{q(q+1)} \end{bmatrix}$ for $\Phi_{qqm} = \bigoplus_{r=q+1}^{Q} I_{N_{rm}} \otimes \pi_{qq}^r$, we structure $V_m^{qq} = \begin{bmatrix} I_{N_{qm}} \otimes \pi_{qq}^q + Z_{-qqm} \Phi_{qqm} Z_{-qqm}^T & Z_{-qqm} \Phi_m^{q(q+1)} Z_m^{(q+1)T} \\ Z_m^{q+1} \Phi_m^{q(q+1)T} Z_{-qqm}^T & V_m^{(q+1)(q+1)} \end{bmatrix}$ in a familiar form (Shin and Raudenbush, 2007) that is uniform for all q and has recursive $V_m^{(q+1)(q+1)} = Z_m^{q+1} \Pi_m^{(q+1)(q+1)} Z_m^{(q+1)T}$.

Shin and Raudenbush (2007) illustrated how to efficiently estimate the two-level model

(7) for $Y_m^1 = [Y_{1m}^T \; Y_{2m}^T]^T$ by ML via the EM algorithm where Y_{1m} and Y_{2m} are vectors of arbitrary length. The key insight of the model (7) is to express the familiar two-level form $Y_m^q = [Y_{qm}^T \; Y_m^{(q+1)T}]^T$ uniformly at each level q and estimate it by the method of Shin and Raudenbush (2007) given computation components at level $q+1$, starting from the highest level $q = Q$ until we estimate the desired model for Y_m^1 with arbitrary Q levels at $q = 1$. The initial step is to estimate the single-level model (7) for $q = Q$, a special case of Shin and Raudenbush's two-level model. We formalize the recursive estimation within each iteration of the EM algorithm after defining notation for estimation below. A major advantage of this approach is that the step-q estimation uses only level-q data given higher-level computation components for efficient computation as will be shown in the following section.

To express the relationship between complete and observed data, let $O_{qm} = \bigoplus_{i=1}^{N_{qm}} O_{qim}$ be a matrix of dummy indicators for observed values in $Y_{qm} = [Y_{q1m}^T \; Y_{q2m}^T \cdots Y_{qN_{qm}m}^T]^T$ so that $Y_{qmobs} = O_{qm}Y_{qm}$ and $Z_{qmobs} = O_{qm}Z_{qm} = [O_{qm} \; Z_{-qqmobs}]$ for $Z_{-qqmobs} = O_{qm}Z_{-qqm}$ (Shin and Raudenbush, 2007). At level Q, $N_{Qm} = 1$ and $O_{Qm} = O_{Q1m}$ for all m. Let $Y_{mobs}^q = O_m^q Y_m^q$ and $Z_{mobs}^q = O_m^q Z_m^q$ for $O_m^q = \bigoplus_{r=q}^{Q} O_{rm}$ to express the observed model (7)

$$Y_{mobs}^q = Z_{mobs}^q b_m^q \sim N(0, V_{mobs}^{qq}), \qquad V_{mobs}^{qq} = O_m^q V_m^{qq} O_m^{qT} = Z_{mobs}^q \Pi_m^{qq} Z_{mobs}^{qT}. \tag{8}$$

Estimation of $\pi = (\pi_1, \pi_2, \cdots, \pi_Q)$ is carried out via the EM algorithm. The complete data (CD) for cluster m may be viewed as b_m^1 given π. Let $\epsilon = (b_1^1, b_2^1, \cdots, b_{N_Q}^1)$ for the entire sample. If we denote $\epsilon_{qim} = \epsilon_q$ in column q of Table 1 for unit i nested within cluster m, the CD log likelihood of π may be expressed as $l(\pi|\epsilon) = \sum_{q=1}^{Q}\sum_{m=1}^{N_Q}\sum_{i=1}^{N_{qm}} lnf(\epsilon_{qim}|\pi)$ for the density $f(\epsilon_{qim}|\pi)$ of level-q unit-specific $\epsilon_{qim} \sim N(0, \pi_q)$. The CD ML estimators are $\hat{\pi}_q = \sum_{m=1}^{N_Q}\sum_{i=1}^{N_{qm}} \epsilon_{qim}\epsilon_{qim}^T/N_q$. The E components are from $b_m^1|Y_{mobs}^1 \sim N(\tilde{b}_m^1, \tilde{\Pi}_m^{11})$ the conventional estimation of which requires inversion of V_{mobs}^{11} that may be extremely high dimensional. The orthogonal decomposition by level of the joint model (1) enables expression of $b_m^q|Y_{mobs}^q \sim N(\tilde{b}_m^q, \tilde{\Pi}_m^{qq})$ that is uniform for all q and based on level-q data only given computation components from higher levels. As will be explained below, this recursive expression yields successive Q-step estimation formulas from $(\tilde{b}_m^Q, \tilde{\Pi}_m^{QQ})$ down to $(\tilde{b}_m^1, \tilde{\Pi}_m^{11})$ for efficient computation of the E step without directly inverting V_{mobs}^{11}.

To estimate fixed effects, let $\mu = [\mu_1 \; \mu_2 \cdots \mu_Q]$ for $\mu_q = X_q\beta_q$ in the model (1) where X_q is a matrix of completely observed covariates having fixed effects β_q. We replace $Y_q = Z_q b_q$ with $d_q = Y_q - X_q\beta_q = Z_q b_q$ in the level-q model (2) to express the stacked-up model (5) as $d^q = Y^q - X^q\beta^q = Z^q b^q$ for $d^q = [d_q^T \; d_{(q+1)}^T \cdots d_Q^T]^T$, $X^q = \bigoplus_{r=q}^{Q} X_r$ and $\beta^q = [\beta_q^T \; \beta_{(q+1)}^T \cdots \beta_Q^T]^T$. Let $d_{qm} = Y_{qm} - X_{qm}\beta_q$ aggregate $d_q = Y_q - X_q\beta_q$ such that the corresponding model (7) has $d_m^q = Y_m^q - X_m^q\beta^q = Z_m^q b_m^q$ for $d_m^q = [d_{qm}^T \; d_{(q+1)m}^T \cdots d_{Qm}^T]^T$ and $X_m^q = \bigoplus_{r=q}^{Q} X_{rm}$. Then, $d_{mobs}^q = O_m^q d_m^q$ and $X_{mobs}^q = O_m^q X_m^q$ so that the observed model (8) is $d_{mobs}^q = Y_{mobs}^q - X_{mobs}^q\beta^q = Z_{mobs}^q b_m^q$. The desired parameters are $\theta = (\pi, \beta)$ for $\beta = \beta^1$. For simplicity of notations, let V_{mobs}^{-q} and $V_{mobs}^{-(q+1)}$ denote the inverses of V_{mobs}^{qq} and $V_{mobs}^{(q+1)(q+1)}$, respectively. Given current β_0, the Fisher scoring equivalent to the Newton-Raphson update is

$$\hat{\beta} \;\; = \;\; \beta_0 + \left(\sum_{m=1}^{N_Q} X_{mobs}^{1T} V_{mobs}^{-1} X_{mobs}^1\right)^{-1} \sum_{m=1}^{N_Q} X_{mobs}^{1T} V_{mobs}^{-1} d_{mobs}^1. \tag{9}$$

The following section describes efficient recursive computation of $\hat{\beta}$ based on $X^{qT}_{mobs}V^{-q}_{mobs}X^{q}_{mobs}$ and $X^{qT}_{mobs}V^{-q}_{mobs}d^{q}_{mobs}$.

Equation (7) expresses a single-level GLM when Z^1_m is an identity matrix and b^1_m has $\Pi^{11}_m = \Pi^{11}$ for all m. The clustering of sample data discussed above gives rise to a Q-level GLM. Next, we show that the aggregated joint model (7) enables us to write down efficient Q-step recursive estimation formulas the qth-step computation of which involves level-q data only and thus is not unduly burdened with respect to the number of Q levels, p variables and random effects.

3.3 Efficient Computation

The conventional E step based on $b^1_m|Y^1_{mobs} \sim N(\tilde{b}^1_m, \tilde{\Pi}^{11}_m)$ requires inversion of V^{11}_{mobs} which may be extremely high dimensional and, thus, take long to compute within each iteration of the EM algorithm. On the other hand, the E step based on Equation (8) produces Q-step estimation formulas the qth step of which is to estimate $b^q_m|Y^q_{mobs}, \theta \sim N(\tilde{b}^q_m, \tilde{\Pi}^{qq}_m)$ where

$$\tilde{b}^q_m = \Pi^{qq}_m A^q_{mobs}, \qquad \tilde{\Pi}^{qq}_m = \Pi^{qq}_m - \Pi^{qq}_m B^q_{mobs} \Pi^{qq}_m \tag{10}$$

for $A^q_{mobs} = Z^{qT}_{mobs}V^{-q}_{mobs}d^q_{mobs}$ and $B^q_{mobs} = Z^{qT}_{mobs}V^{-q}_{mobs}Z^q_{mobs}$. The key advantages of the E step via Equations (10) are that $\tilde{b}^q_m$ and $\tilde{\Pi}^{qq}_m$ stay uniform for all q; that computation of $\tilde{b}^q_m$ and $\tilde{\Pi}^{qq}_m$ uses level-q data only, given A^{q+1}_{mobs}, B^{q+1}_{mobs} and θ; that the expressions (10) enable efficient computation of $\tilde{b}^1_m$ and $\tilde{\Pi}^{11}_m$ via computation of recursive components A^q_{mobs} and B^q_{mobs}; and that the E step does not require direct inversion of V^{11}_{mobs}. Estimation of $\tilde{b}^1_m$ and $\tilde{\Pi}^{11}_m$ starts at the highest level $q = Q$ with initial components

$$A^Q_{mobs} = O^T_{Qm}V^{-Q}_{mobs}d_{Qmobs}, \qquad B^Q_{mobs} = O^T_{Qm}V^{-Q}_{mobs}O_{Qm} \tag{11}$$

for $V^{QQ}_{mobs} = \pi^Q_{QQm} = O_{Qm}\pi^Q_{QQ}O^T_{Qm}$, computes A^q_{mobs} and B^q_{mobs} using level-q data only, given A^{q+1}_{mobs} and B^{q+1}_{mobs} at step q, and finally evaluates Equations (10) given A^1_{mobs} and B^1_{mobs} at $q = 1$ within each iteration of the EM algorithm.

To formulate the recursive computation, let

$$\begin{aligned}
\tilde{\epsilon}_{-qqm} &= E(\epsilon_{-qqm}|Y^q_{mobs}) = \Delta^{-1}_{qm}(Z_{-qqmobs}\psi^{-1}_{qm}d_{qmobs} + \Omega^{-1}_{qm}\tilde{\tilde{\epsilon}}_{-qqm}), \\
V^{-q11}_{mobs} &= \psi^{-1}_{qm} - \psi^{-1}_{qm}Z_{-qqmobs}\Delta^{-1}_{qm}Z^T_{-qqmobs}\psi^{-1}_{qm}, \\
V^{-q}_{mobs} &= \left[\begin{array}{cc} V^{-q11}_{mobs} & V^{-q12}_{mobs} \\ V^{-q21}_{mobs} & V^{-q22}_{mobs} \end{array} \right]
\end{aligned} \tag{12}$$

where $\Delta_{qm} = Z^T_{-qqmobs}\psi^{-1}_{qm}Z_{-qqmobs} + \Omega^{-1}_{qm}$, $\psi_{qm} = \bigoplus_{i=1}^{N_{qm}} \pi^q_{qqim}$, $\Omega_{qm} = var(\epsilon_{-qqm}|Y^{q+1}_{mobs}) = \Phi_{qqm} - \Phi^{q(q+1)}_m B^{q+1}_{mobs}\Phi^{q(q+1)T}_m$ and $\tilde{\tilde{\epsilon}}_{-qqm} = E(\epsilon_{-qqm}|Y^{q+1}_{mobs}) = \Phi^{q(q+1)}_m A^{q+1}_{mobs}$ for $\pi^q_{qqim} = O_{qim}\pi^q_{qq}O^T_{qim}$. Note that computation of $\tilde{\epsilon}_{-qqm}$ and V^{-q11}_{mobs} requires level-q $Z_{-qqmobs}$ and d_{qmobs} only, given A^{q+1}_{mobs}, B^{q+1}_{mobs} and θ. The following result shows that A^q_{mobs} and B^q_{mobs} depend on level-q data only, given higher-level components A^{q+1}_{mobs} and B^{q+1}_{mobs}. See the Appendix for proofs.

Proposition 3.1 *$Z^{qT}_{mobs}V^{-q}_{mobs}d^{q}_{mobs}$ and $Z^{qT}_{mobs}V^{-q}_{mobs}Z^{q}_{mobs}$ depend on level-q data Y_{qmobs}, X_{qmobs} and $Z_{-qqmobs}$ only, given $Z^{(q+1)T}_{mobs}V^{-(q+1)}_{mobs}d^{q+1}_{mobs}$, $Z^{(q+1)T}_{mobs}V^{-(q+1)}_{mobs}Z^{q+1}_{mobs}$ and θ for all $q < Q$.*

Proposition 3.1 enables a Q-step recursive computation of $\tilde{b}^1_m$ and $\tilde{\Pi}^{11}_m$ the qth step of which involves level-q data only without directly inverting V^{11}_{mobs}.

Theorem 3.2 *$\tilde{b}^1_m$ and $\tilde{\Pi}^{11}_m$ can be computed by a Q-step recursive procedure from level Q down to level 1 given θ where step q involves level-q data only.*

The E step for $Q = 3$, for example, computes $A^3_{mobs} = O^T_{3m}\pi^{-3}_{33m}d_{3mobs}$ and $B^3_{mobs} = O^T_{3m}\pi^{-3}_{33m}O_{3m}$ in Equations (11) initially for the inverse π^{-3}_{33m} of π^3_{33m}; A^2_{mobs} and B^2_{mobs} in Equations (32) and (33) given A^3_{mobs} and B^3_{mobs} at $q = 2$; A^1_{mobs} and B^1_{mobs} in Equations (32) and (33) given A^2_{mobs} and A^2_{mobs} to finally yield $\tilde{b}^1_m$ and $\tilde{\Pi}^{11}_m$ in Equations (10) at $q = 1$.

Fisher scoring on β may also be based on recursive computation of A^q_{mobs}, B^q_{mobs},

$$F^q_{mobs} = X^{qT}_{mobs}V^{-q}_{mobs}d^q_{mobs}, \quad G^q_{mobs} = X^{qT}_{mobs}V^{-q}_{mobs}X^q_{mobs}, \quad H^q_{mobs} = Z^{qT}_{mobs}V^{-q}_{mobs}X^q_{mobs}. \quad (13)$$

The following result shows that F^q_{mobs}, G^q_{mobs} and H^q_{mobs} depend on level-q data only, given A^{q+1}_{mobs}, B^{q+1}_{mobs}, F^{q+1}_{mobs}, G^{q+1}_{mobs}, H^{q+1}_{mobs} and θ.

Proposition 3.3 *$X^{qT}_{mobs}V^{-q}_{mobs}d^q_{mobs}$, $X^{qT}_{mobs}V^{-q}_{mobs}X^q_{mobs}$ and $Z^{qT}_{mobs}V^{-q}_{mobs}X^q_{mobs}$ depend on level-q data Y_{qmobs}, X_{qmobs} and $Z_{-qqmobs}$ only, given $Z^{(q+1)T}_{mobs}V^{-(q+1)}_{mobs}d^{q+1}_{mobs}$, $Z^{(q+1)T}_{mobs}V^{-(q+1)}_{mobs}Z^{q+1}_{mobs}$, $X^{(q+1)T}_{mobs}V^{-(q+1)}_{mobs}d^{q+1}_{mobs}$, $X^{(q+1)T}_{mobs}V^{-(q+1)}_{mobs}X^{q+1}_{mobs}$, $Z^{(q+1)T}_{mobs}V^{-(q+1)}_{mobs}X^{q+1}_{mobs}$ and θ for all $q < Q$.*

Propositions 3.1 and 3.3 enable $\hat{\beta}$ to be computed recursively.

Theorem 3.4 *$X^{1T}_{mobs}V^{-1}_{mobs}X^1_{mobs}$ and $X^{1T}_{mobs}V^{-1}_{mobs}d^1_{mobs}$ can be computed by a Q-step recursive procedure from level Q down to level 1 given θ where step q involves level-q data only.*

The Fisher scoring on β for $Q = 3$, for example, computes $A^3_{mobs} = O^T_{3m}\pi^{-3}_{33m}d_{3mobs}$, $B^3_{mobs} = O^T_{3m}\pi^{-3}_{33m}O_{3m}$, $F^3_{mobs} = X^T_{3mobs}\pi^{-3}_{33m}d_{3mobs}$, $G^3_{mobs} = X^T_{3mobs}\pi^{-3}_{33m}X_{3mobs}$ and $H^3_{mobs} = O^T_{3mobs}\pi^{-3}_{33m}X_{3mobs}$ initially; A^2_{mobs}, B^2_{mobs}, F^2_{mobs}, G^2_{mobs} and H^2_{mobs} in Equations (32) to (36) from level-2 data, given A^3_{mobs}, B^3_{mobs}, F^3_{mobs}, G^3_{mobs} and H^3_{mobs} at $q = 2$; F^1_{mobs} and G^1_{mobs} in Equations (34) and (35) from level-1 data given A^2_{mobs}, B^2_{mobs}, F^2_{mobs}, G^2_{mobs} and H^2_{mobs} to finally yield $\hat{\beta}$ in Equation (9) at $q = 1$.

The recursive estimation efficiently handles missing data one level at a time. Consequently, the computation will be efficient given a number of variables subject to missingness at higher levels. In that case, the observed joint model (8) yields recursive computation that is not excessively burdened with respect to Q, p and the number of random effects. On the other hand, given missing data at level 1 only, this approach amounts to the conventional EM algorithm. The inverse of the Fisher information matrix yields $var(\hat{\theta})$.

Multiple imputation is based on $Y^1_m|Y^1_{mobs}, \hat{\theta} \sim N\left(X^1_m\beta^1 + Z^1_m\tilde{b}^1_m, Z^1_m\tilde{\Pi}^{11}_m Z^{1T}_m\right)$ given the ML $\hat{\theta}$. Let v_{ii} and v_{ij} be variances of $log(\pi_{ii})$ and $log\frac{1+\rho_{ij}}{1-\rho_{ij}}$ where $i \neq j$ and $\rho_{ij} = \frac{\pi_{ij}}{\sqrt{\pi_{ii}\pi_{jj}}}$ for diagonal π_{ii} and off-diagonal π_{ij} in $(\pi_1, \pi_2, \cdots, \pi_Q)$. Then, $N[log(\hat{\pi}_{ii}), \hat{v}_{ii}]$ and $N\left(log\frac{1+\hat{\rho}_{ij}}{1-\hat{\rho}_{ij}}, \hat{v}_{ij}\right)$ estimated by ML imply the joint distribution of a vector, ϕ_q, of distinct $log(\pi_{ii})$ and $log\frac{1+\rho_{ij}}{1-\rho_{ij}}$ from π_q. Let the distributions of ϕ_q and β be $N\left[\hat{\phi}_q, var(\hat{\phi}_q)\right]$ and $N[\hat{\beta}, var(\hat{\beta})]$ estimated by

ML, respectively. To propagate uncertainty in estimation of θ for proper imputation (Little and Rubin, 2002), we randomly generate β from $N[\hat{\beta}, var(\hat{\beta})]$ and π_q from $N\left[\hat{\phi}_q, var(\hat{\phi}_q)\right]$ for all q, and then impute missing data given the θ for each imputation (Shin and Raudenbush, 2007).

Thus far, we have focused on estimating the joint model (1) for variables subject to missingness given completely observed covariates. However, our goal in this paper is to estimate a general Q-level hierarchial GLM for a univariate response conditional on covariates where the covariates as well as the response may have ignorable missing data at any of the levels. To efficiently estimate the GLM, we have to reparameterize it in the form of the joint model (1). The next section will introduce the desired GLM, clarify its relationship with the joint model, and describe methods to efficiently estimate the GLM via the joint model.

4 Hierarchical General Linear Model

The aim of this article is to estimate a Q-level hierarchial GLM that is a special case of the joint model (1) in which a univariate response is defined at the lowest level of aggregation. We show that the joint model overidentifies the desired model, in general, and describe how to constrain the joint model to be a one-to-one transformation of the GLM. Without the one-to-one correspondence, the estimated GLM via MLE *on* Y_{obs} may be substantially biased (Shin and Raudenbush, 2007). For simplicity of explication, we first consider the desired GLM where all covariates having fixed effects are subject to missingness. This consideration is without loss of generality because completely observed covariates having fixed effects do not affect the constraints on the joint model. After explaining the needed constraints for MLE *on* Y_{obs}, we consider a more realistic GLM having both covariates subject to missingness and covariates completely observed.

We write the general Q-level hierarchial GLM

$$R = C^T\gamma + D^T e, \qquad e \sim N(0, \tau) \tag{14}$$

for $C = [A^T \; Y_2^T \cdots Y_Q^T]^T$ having fixed effects $\gamma = [\gamma_1^T \; \gamma_2^T \cdots \gamma_Q^T]^T$, $D = [D_1^T \; D_2^T \cdots D_Q^T]^T$ having random effects $e = [e_1^T \; e_2^T \cdots e_Q^T]^T$, and $\tau = \bigoplus_{r=1}^{Q} \tau_r$ where A and Y_r are vectors of $p_1 - 1$ level-1 and p_r level-r covariates having fixed effects γ_1 and γ_r, respectively, D_r is a vector of p_{Dr} covariates having level-r unit-specific random effects $e_r \sim N(0, \tau_r)$ independent across levels and $p_D = \sum_{r=1}^{Q} p_{Dr}$. Both R and C are subject to missingness while D is known. We assume $D_1 = 1$ and that D_r carries an intercept as many applications do, although it is not required to have one.

The aim of this article is to efficiently estimate the Q-level hierarchical GLM (14) given incomplete data. To do so, we must reparameterize the equation (14) in the form of the joint distribution (1) of all variables subject to missingness - including the response and covariates at any level given D. We define the first element of Y_1 as the response R and the remaining elements of Y_1 as covariates A to partition $Y_1 = [R \; A^T]^T$ and decompose Y into the response and covariates $Y = [R \; C^T]^T$ where we decompose $R = \sum_{r=1}^{Q} D_r^T \epsilon_{rR}$ and $C = [A^T \; Y_2^T \cdots Y_Q^T]^T = [\sum_{r=1}^{Q} \epsilon_{rA}^T \; \sum_{r=2}^{Q} \epsilon_{r2}^T \cdots \epsilon_{QQ}^T]^T$ orthogonally by level. Then, Equation

(14) is a special case of the model (1) for $\mu = 0$ where $Y_1 = [R \; A^T]^T$ implies partitioning

$$Z_{r1} = \begin{bmatrix} D_r^T & 0 \\ 0 & I_{p_1-1} \end{bmatrix}, \; \epsilon_{r1} = \begin{bmatrix} \epsilon_{rR} \\ \epsilon_{rA} \end{bmatrix}, \; \pi_{11}^r = \begin{bmatrix} \pi_{RR}^r & \pi_{RA}^r \\ \pi_{AR}^r & \pi_{AA}^r \end{bmatrix}, \; \pi_{1q}^r = \begin{bmatrix} \pi_{Rq}^r \\ \pi_{Aq}^r \end{bmatrix} \tag{15}$$

and C implies $Z_{rq} = I_{p_q}$ for $q > 1$. Then, $V = \begin{bmatrix} var(R) & cov(R,C) \\ cov(C,R) & var(C) \end{bmatrix} = \begin{bmatrix} V_{RR} & V_{RC} \\ V_{RC}^T & V_{CC} \end{bmatrix}$ for

$$\begin{aligned} V_{RR} &= \sum_{r=1}^{Q} D_r^T \pi_{RR}^r D_r, \;\; V_{RC} = \left[\sum_{r=1}^{Q} D_r^T \pi_{RA}^r \; \sum_{r=2}^{Q} D_r^T \pi_{R2}^r \cdots D_Q^T \pi_{RQ}^Q \right], & (16) \\ V_{CC} &= \begin{bmatrix} \pi_{AA}^1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} + \begin{bmatrix} \pi_{AA}^2 & \pi_{A2}^2 & \cdots & 0 \\ \pi_{2A}^2 & \pi_{22}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} + \cdots + \begin{bmatrix} \pi_{AA}^Q & \pi_{A2}^Q & \cdots & \pi_{AQ}^Q \\ \pi_{2A}^Q & \pi_{22}^Q & \cdots & \pi_{2Q}^Q \\ \vdots & \vdots & \ddots & \vdots \\ \pi_{QA}^Q & \pi_{Q2}^Q & \cdots & \pi_{QQ}^Q \end{bmatrix}. \end{aligned}$$

We refer to $Y = [\sum_{r=1}^{Q} D_r^T \epsilon_{rR} \; \sum_{r=1}^{Q} \epsilon_{rA}^T \; \sum_{r=2}^{Q} \epsilon_{r2}^T \cdots \epsilon_{QQ}^T]^T$ as the form of the joint model (1) for efficient estimation of the GLM (14) in this article. Then, the GLM (14) implies

$$var(R) = \gamma^T V_{CC} \gamma + D^T \tau D \text{ and } cov(R,C) = \gamma^T V_{CC}.$$

When $Q = 1$, the reparameterization is one-to-one between Equations (1) and (14) and no difficulties arise in computation and interpretation as illustrated in Section 4.1. However, when $Q > 1$, we find that the reparameterization required to equate the conditional model (14) to the corresponding joint model (1) can be quite challenging. Without imposing constraints, the joint model will over-identify the conditional model (14). We can readily comprehend this problem in the case of two-level models shown in Sections 4.2 and 4.3. We then generalize our approach in the subsequent section. We see that the problem of over-identification can become severe as covariates, levels and random coefficients are added to the model.

4.1 Single-Level Model

Equation (14) for $C = A$, $\gamma = \gamma_1$, $D = 1$ and $e = e_1$ expresses the conventional ordinary least squares (OLS) regression model as the conditional distribution of R given covariates A. Efficient estimation of the model from ignorable missing data (Rubin, 1976, Little and Rubin, 2002) is straightforward when we estimate the corresponding joint model (1)

$$\begin{bmatrix} R \\ A \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} V_{RR} & V_{RC} \\ V_{RC}^T & V_{CC} \end{bmatrix} \right) \tag{17}$$

for $V_{RR} = \pi_{RR}^1$, $V_{RC} = \pi_{RA}^1$ and $V_{CC} = \pi_{AA}^1$. The OLS model (14) implies $V_{RR} = \gamma_1^T V_{CC} \gamma_1 + \tau_1$ and $V_{RC} = \gamma_1^T V_{CC}$ to yield the one-to-one transformations $\gamma_1^T = V_{RC} V_{CC}^{-1}$ and $\tau_1 = V_{RR} - \gamma_1^T V_{CC} \gamma_1$. That is, the p_1 parameters in the OLS model (14) and the $(p_1 - 1)p_1/2$ variance and covariance components in V_{CC} are one-to-one functions of the $p_1(p_1 + 1)/2$ parameters in the joint model (17). Equivalently, the p_1 parameters in the OLS model are one-to-one functions of (V_{RR}, V_{RC}) without redundantly counting the number of parameters in V_{CC}.

The general forms of the joint model (1) implied by the equations (16) and the conditional model (14) remain intact when we consider the hierarchical linear model with arbitrary Q levels. A central concern of interest to this paper is that, when we move beyond the single level case for $Q = 1$, the desired model (14) is not a one-to-one transformation of the joint model. To see how this works, we consider two-level data where level-1 units (e.g. students) are nested within level-2 units (e.g. schools) before we consider arbitrary Q levels. We shall consider the cases of the two-level model with a random intercept and the two-level model with random coefficients.

4.2 Random-Intercept Model

A comparatively simple two-level hierarchical linear model with a random intercept is of form

$$R \;=\; A^T\gamma_1 + Y_2^T\gamma_2 + e_1 + e_2 \sim N(A^T\gamma_1 + Y_2^T\gamma_2, \tau_1 + \tau_2), \tag{18}$$

a special case of model (14) with $C = [A^T\ Y_2^T]^T$, $\gamma = [\gamma_1^T\ \gamma_2^T]^T$, $D = [1\ 1]^T$, $e = [e_1\ e_2]^T$ and $\tau = \bigoplus_{q=1}^2 \tau_q$. The corresponding joint model (1) is

$$\begin{bmatrix} R \\ A \\ Y_2 \end{bmatrix} = \begin{bmatrix} \epsilon_{1R} \\ \epsilon_{1A} \\ 0 \end{bmatrix} + \begin{bmatrix} \epsilon_{2R} \\ \epsilon_{2A} \\ \epsilon_{22} \end{bmatrix} \sim N\left(0, \begin{bmatrix} \pi_{RR}^1 & \pi_{RA}^1 & 0 \\ \pi_{AR}^1 & \pi_{AA}^1 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} \pi_{RR}^2 & \pi_{RA}^2 & \pi_{R2}^2 \\ \pi_{AR}^2 & \pi_{AA}^2 & \pi_{A2}^2 \\ \pi_{2R}^2 & \pi_{2A}^2 & \pi_{22}^2 \end{bmatrix}\right) \tag{19}$$

where $V_{RR} = \pi_{RR}^1 + \pi_{RR}^2$ and $V_{RC} = [\pi_{RA}^1 + \pi_{RA}^2\ \pi_{R2}^2]$ and $V_{CC} = \begin{bmatrix} \pi_{AA}^1 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} \pi_{AA}^2 & \pi_{A2}^2 \\ \pi_{2A}^2 & \pi_{22}^2 \end{bmatrix}$. We can see now that the desired model (18) constrains the joint model (19) by $V_{RR} = \gamma^T V_{CC}\gamma + \tau_1 + \tau_2$ and $V_{RC} = \gamma^T V_{CC}$ such that

$$\pi_{RR}^1 = [\gamma_1^T\ \gamma_2^T] \begin{bmatrix} \pi_{AA}^1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} + \tau_1, \qquad [\pi_{RA}^1\ 0] = [\gamma_1^T\ \gamma_2^T] \begin{bmatrix} \pi_{AA}^1 & 0 \\ 0 & 0 \end{bmatrix}, \tag{20}$$

$$\pi_{RR}^2 = [\gamma_1^T\ \gamma_2^T] \begin{bmatrix} \pi_{AA}^2 & \pi_{A2}^2 \\ \pi_{2A}^2 & \pi_{22}^2 \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} + \tau_2, \qquad [\pi_{RA}^2\ \pi_{R2}^2] = [\gamma_1^T\ \gamma_2^T] \begin{bmatrix} \pi_{AA}^2 & \pi_{A2}^2 \\ \pi_{2A}^2 & \pi_{22}^2 \end{bmatrix}. \tag{21}$$

To see how many constraints the desired model (18) has placed on the joint model (19), the constrained model (18) identifies $p_1 + p_2 + 1$ parameters while the unconstrained joint model (19) identifies $2p_1 + p_2$ parameters in (V_{RR}, V_{RC}). Therefore, the constrained model (18) has $p_1 - 1$ fewer parameters than does the unconstrained joint model (19). The key constraints occur in the variances and covariances (20) and (21) where the association γ_1 between R and A is constrained to be the same at both levels. An alternative form of the unconstrained model (19) replaces γ_1 with γ_{11} in the equations (20) and γ_1 with γ_{12} in the equations (21). This would allow the association between R and A to be different at the two levels, inducing what is known in the social science and public health applications as a contextual effects model (Shin and Raudenbush, 2010). The constraints (20) and (21) impose $\gamma_{12} - \gamma_{11} = 0$, that is, no contextual effects.

4.3 Random-Coefficients Model

As the number p_D of random coefficients increases, the number of potentially extraneous parameters generated will increase non-linearly if no constraints are imposed. To show how aggravated the over-identification can become, consider a random coefficients model that adds level-1 covariates E_2 having random coefficients to the model (18)

$$R \;=\; A^T\gamma_1 + Y_2^T\gamma_2 + e_1 + D_2^T e_2 \sim N(A^T\gamma_1 + Y_2^T\gamma_2, \tau_1 + D_2^T\tau_2 D_2), \qquad (22)$$

another special case of model (14) for $C = [A^T \; Y_2^T]^T$, $\gamma = [\gamma_1^T \; \gamma_2^T]^T$, $D = [D_1 \; D_2^T]^T$, $e = [e_1 \; e_2^T]^T$ and $\tau = \bigoplus_{q=1}^2 \tau_q$ where $D_1 = 1$, $D_2 = \begin{bmatrix} 1 \\ E_2 \end{bmatrix}$, $e_2 = \begin{bmatrix} e_{20} \\ e_{21} \end{bmatrix}$, $\tau_2 = \begin{bmatrix} \tau_{200} & \tau_{201} \\ \tau_{210} & \tau_{211} \end{bmatrix}$ and $p_D = 1 + p_{D2}$. The corresponding joint model (1) is

$$\begin{bmatrix} R \\ A \\ Y_2 \end{bmatrix} = \begin{bmatrix} \epsilon_{1R} \\ \epsilon_{1A} \\ 0 \end{bmatrix} + \begin{bmatrix} D_2^T\epsilon_{2R} \\ \epsilon_{2A} \\ \epsilon_{22} \end{bmatrix} \sim N\left(0, \begin{bmatrix} \pi^1_{RR} & \pi^1_{RA} & 0 \\ \pi^1_{AR} & \pi^1_{AA} & 0 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} D_2^T\pi^2_{RR}D_2 & D_2^T\pi^2_{RA} & D_2^T\pi^2_{R2} \\ \pi^2_{AR}D_2 & \pi^2_{AA} & \pi^2_{A2} \\ \pi^2_{2R}D_2 & \pi^2_{2A} & \pi^2_{22} \end{bmatrix}\right) (23)$$

for $\epsilon_{2R} = \begin{bmatrix} \epsilon_{2R0} \\ \epsilon_{2R1} \end{bmatrix}$, $\pi^2_{RR} = \begin{bmatrix} \pi^2_{R0R0} & \pi^2_{R0R1} \\ \pi^2_{R1R0} & \pi^2_{R1R1} \end{bmatrix}$, $\pi^2_{RA} = \begin{bmatrix} \pi^2_{R0A} \\ \pi^2_{R1A} \end{bmatrix}$ and $\pi^2_{R2} = \begin{bmatrix} \pi^2_{R02} \\ \pi^2_{R12} \end{bmatrix}$. Note that $V_{RR} = \pi^1_{RR} + D_2^T\pi^2_{RR}D_2$, $V_{RC} = [\pi^1_{RA} + D_2^T\pi^2_{RA} \;\; D_2^T\pi^2_{R2}]$ and V_{CC} as in the model (19). The desired model (22) implies constraining the joint model (23) by $V_{RR} = \gamma^T V_{CC}\gamma + \tau_1 + D_2^T\tau_2 D_2$ and $V_{RC} = \gamma^T V_{CC}$ such that

$$\pi^1_{RR} = [\gamma_1^T \; \gamma_2^T] \begin{bmatrix} \pi^1_{AA} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} + \tau_1, \quad [\pi^1_{RA} \; 0] = [\gamma_1^T \; \gamma_2^T] \begin{bmatrix} \pi^1_{AA} & 0 \\ 0 & 0 \end{bmatrix}, \qquad (24)$$

$$\pi^2_{R0R0} = [\gamma_1^T \; \gamma_2^T] \begin{bmatrix} \pi^2_{AA} & \pi^2_{A2} \\ \pi^2_{2A} & \pi^2_{22} \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} + \tau_{200}, \; \left[\pi^2_{R0A} \; \pi^2_{R02}\right] = [\gamma_1^T \; \gamma_2^T] \begin{bmatrix} \pi^2_{AA} & \pi^2_{A2} \\ \pi^2_{2A} & \pi^2_{22} \end{bmatrix}, \qquad (25)$$

$$\pi^2_{R0R1} = \tau_{201}, \qquad \pi^2_{R1R1} = \tau_{211}, \qquad \pi^2_{R1A} = 0, \qquad \pi^2_{R12} = 0.$$

To see how many constraints the desired model (22) has placed on the joint model (23), the constrained model (22) identifies $p_1 + p_2 + p_{D2}(p_{D2}+1)/2$ parameters while the unconstrained joint model (23) has $p_1 + p_{D2}(p_{D2}+1)/2 + p_{D2}(p_1 + p_2 - 1)$ components in (V_{RR}, V_{RC}). Therefore, the model (22) has $p_{D2}(p_1 + p_2 - 1) - p_2$ fewer parameters than does the unconstrained joint model (23). Again, the key constraints occur in the variances and covariances (24) and (25) where not only is the association γ_1 between R and A constrained to be the same at both levels, but the covariance components π^2_{R1A} and π^2_{R12} that yield extraneous interaction effects between E_2 and C are also set to zero. That is, the desired model (22) has no contextual effects of A and no interaction effects between E_2 and C. Next, we extend the model (14) to an arbitrary number of Q levels.

4.4 Q-Level Model

We now focus on how to efficiently estimate the hierarchical GLM (14) for the arbitrary number of Q levels. Unlike the single-level case, however, the joint distribution (1) over-identifies the desired model (14). This over-identification poses a major computational challenge as it represents the components of $cov(R, C) = V_{RC}$ that are extraneous for subsequent analysis and that rapidly multiply as Q, p and p_D increase. The consequence is that estimation of the over-identified hierarchical model (14) may produce substantially biased inferences in the case of $MLE\ on\ Y_{obs}$ or computational problems in the case of $MLE\ on\ Y^{mi}$.

To show the over-identification explicitly, we reexpress all random effects of the joint model (1) in Table 1 according to the decomposition $Y = [R\ C^T]^T$ as listed in Table 2. Column q lists level-q unit-specific random effects $\epsilon_q \sim N(0, \pi_q)$ that may now be partitioned as

Table 2: All random components of Y in Equation (1).

	1	2	3	$\cdots$	Q
R	ϵ_{1R}	ϵ_{2R}	ϵ_{3R}	$\cdots$	ϵ_{QR}
A	ϵ_{1A}	ϵ_{2A}	ϵ_{3A}	$\cdots$	ϵ_{QA}
Y_2		ϵ_{22}	ϵ_{32}	$\cdots$	ϵ_{Q2}
Y_3			ϵ_{33}	$\cdots$	ϵ_{Q3}
$\vdots$				$\ddots$	$\vdots$
Y_Q					ϵ_{QQ}

$$\epsilon_q = \begin{bmatrix} \epsilon_{qR} \\ \epsilon_{qC} \end{bmatrix}, \quad \pi_q = \begin{bmatrix} \pi^q_{RR} & \pi^q_{RC} \\ \pi^q_{CR} & \pi^q_{CC} \end{bmatrix} \quad \text{for } \epsilon_{qC} = [\epsilon^T_{qA}\ \epsilon^T_{q2} \cdots \epsilon^T_{qq}]^T.$$

The random effects $(\epsilon_{qR}, \epsilon_{qC})$ and their variances and covariances $(\pi^q_{RR}, \pi^q_{CC}, \pi^q_{RC})$ are useful for explaining the over-identification problem and the constraints. Notice that the level-q unit-specific ϵ_q generates $p_{Dq}(\sum_{r=1}^q p_r - 1)$ covariance components in π^q_{RC}. For $Q = 3$ with three columns, for example, columns 1, 2 and 3 generate $p_{D1}(p_1 - 1)$, $p_{D2}(p_1 + p_2 - 1)$ and $p_{D3}(p_1 + p_2 + p_3 - 1)$ components in $cov(\epsilon_{1R}, \epsilon_{1C}) = \pi^1_{RC} = \pi^1_{RA}$, $cov(\epsilon_{2R}, \epsilon_{2C}) = \pi^2_{RC} = [\pi^2_{RA}\ \pi^2_{R2}]$ and $cov(\epsilon_{3R}, \epsilon_{3C}) = \pi^3_{RC} = [\pi^3_{RA}\ \pi^3_{R2}\ \pi^3_{R3}]$ at levels 1, 2 and 3, respectively, for $\epsilon_{1C} = \epsilon_{1A}$, $\epsilon_{2C} = [\epsilon^T_{2A}\ \epsilon^T_{22}]^T$ and $\epsilon_{3C} = [\epsilon^T_{3A}\ \epsilon^T_{32}\ \epsilon^T_{33}]^T$. Overall, the random effects of the joint model (1) produce $\sum_{q=1}^Q p_{Dq} \sum_{r=1}^q p_r - p_D$ covariance components between R and C while the desired model (14) implies $p - 1$ elements in γ. Consequently, the potential for severe over-identification exists if no constraints are imposed on Equation (1).

A key task, then, is to formulate a general approach to imposing constraints, one that applies to any value of Q and any number of covariates. The following theorem shows a conditional model the joint model (1) identifies so that constraining the joint model amounts to constraining the conditional model. Let $\epsilon_C = (\epsilon_{1C}, \epsilon_{2C}, \cdots, \epsilon_{QC})$ and π^{-q}_{CC} be the inverse of π^q_{CC}.

Theorem 4.1 *Joint model (1) represents $\sum_{q=1}^{Q} p_{Dq} \sum_{r=1}^{q} p_r - p_D$ covariance components between R and C and is a one-to-one transformation of $R|\epsilon_C, D$*

$$R \;=\; \sum_{q=1}^{Q} \left(D_q^T \Gamma_q \epsilon_{qC} + D_q^T \delta_q \right), \quad \delta_q \sim N(0, \pi_{R|C}^q) \tag{26}$$

for $\Gamma_q = \pi_{RC}^q \pi_{CC}^{-q}$ and $\pi_{R|C}^q = \pi_{RR}^q - \Gamma_q \pi_{CC}^q \Gamma_q^T$.

For each covariate in Y_q, the conditional model (26) expresses the association between the covariate and R to be distinct at each level $s \geq q$ while the desired model (14) represents a single effect of the covariate on R. Consequently, the joint model (1) produces $\sum_{q=1}^{Q} p_{Dq} \sum_{r=1}^{q} p_r - p_D - (p-1)$ parameters extraneous for subsequent analysis. The extraneous parameters, representing the contextual effects of C and the interaction effects between D and ϵ_C, rapidly multiply as Q, p and p_D grow. Let $D_q = [1 \; E_q^T]^T$ and $\epsilon_{qR} = [\epsilon_{qR0} \; \epsilon_{qR1}^T]^T$ so that $\pi_{RC}^q = cov(\epsilon_{qR}, \epsilon_{qC}) = [\pi_{R0C}^{qT} \; \pi_{R1C}^{qT}]^T$. The following corollary to Theorem 4.1 establishes one-to-one correspondence between a general contextual effects model and a constrained joint model (1) where each level-q covariate has a distinct effect at every level $s \geq q$ without involving interaction effects.

Corollary 4.2 *Joint model (1) under constraints*

$$\pi_{R1C}^q = 0, \qquad \forall q \tag{27}$$

identifies a general contextual effects model given ϵ_C and D

$$R = \sum_{q=1}^{Q} \left(\gamma_q^{*T} \epsilon_{qC} + D_q^T \delta_q \right), \quad \delta_q \sim N(0, \pi_{R|C}^q) \tag{28}$$

for $\gamma_q^* = [\gamma_{1q}^T \; \gamma_{2q}^T \cdots \gamma_{qq}^T]^T$.

Equation (28) is nested within Model (26) for $\Gamma_q = \begin{bmatrix} \gamma_q^{*T} \\ 0 \end{bmatrix} = \begin{bmatrix} \pi_{R0C}^q \\ 0 \end{bmatrix} \pi_{CC}^{-q}$. We define the contextual effects of covariates in Y_q at level $s > q$ as $\gamma_{qs} - \gamma_{q(s-1)}$ (Shin and Raudenbush, 2010). Corollary 4.2 may involve γ_q^* expressing constraints of different forms. For example, if it is desirable for each covariate in A to have a single effect on R in the model (28), then the corollary would constrain $\gamma_q^* = [\gamma_1^T \; \gamma_{2q}^T \cdots \gamma_{qq}^T]^T$ for all q in addition to the constraints (27). For another example, if the contextual effects of A are desired at level 2 but no other levels in the model (28), then the additional constraints would be $\gamma_1^* = \gamma_{11}$ and $\gamma_q^* = [\gamma_1^T \; \gamma_{2q}^T \cdots \gamma_{qq}^T]^T$ for $q > 1$. Shin and Raudenbush (2007) imposed $\pi_{R1C}^q = 0$ and a single effect of each covariate on the joint model (1) to identify the desired model (14) for $Q = 2$. The following corollary to Theorem 4.1 establishes the one-to-one correspondence between the model (14) and a constrained joint model (1).

Corollary 4.3 *Joint model (1) under constraints*

$$\pi_{R1C}^q = 0 \quad and \quad \pi_{R0C}^q = [\gamma_1^T \cdots \gamma_q^T] \pi_{CC}^q, \quad \forall q \tag{29}$$

identifies hierarchical model (14).

Model (14) under the constraints (29) is equivalent to model (28) for $\gamma_q^* = [\gamma_1^T \cdots \gamma_q^T]^T$ that implies $\gamma_1^T \sum_{r=1}^{Q} \epsilon_{rA} = \gamma_1^T A$, $\gamma_s^T \sum_{r=s}^{Q} \epsilon_{rs} = \gamma_s^T Y_s$ and $\tau_q = \pi_{R|C}^q$ for all $s > 1$ and all q.

Given Q-level incomplete data, Equation (1) under constraints (29) identifies hierarchical model (14) while the joint model under partial constraints Γ_q such as constraints (27) may identify desired contextual effects of C or interaction effects between D and ϵ_C (Shin and Raudenbush, 2010). All these applications may be carried out via MLE *on* Y_{obs}, MLE *on* Y^{mi} or a hybrid method of imputation following estimation of the constrained joint model. The choice will depend on computational feasibility and the goal of the application. Given an analyst's model (14), MLE *on* Y_{obs} constrains the joint model (1) to just identify the analyst's model while MLE *on* Y^{mi} is more generally applicable by enabling the data analyst to explore, in addition, contextual effects of C and interaction effects involving D. Consequently, MLE *on* Y_{obs} is tailored to estimation of the analyst's model whereas MLE *on* Y^{mi} estimates an overidentified joint model so that it enables the data analyst to explore multiple hierarchical models for correct specification of the analyst's model. When MLE *on* Y^{mi} is desired, but produces an unconstrained joint model (1) that is extremely high dimensional and thus difficult to estimate well, the hybrid method enables estimation of fewer parameters and thus reduces computational burden in estimation by imposing partial constraints such as Equations (27) on the joint model.

Now, we consider a more general model (14)

$$R = C^T\gamma + W^T\gamma_w + D^Te, \qquad e \sim N(0, \tau) \tag{30}$$

for known covariates $W = [W_1^T \; W_2^T \cdots W_Q^T]^T$ having fixed effects $\gamma_w = [\gamma_{w1}^T \; \gamma_{w2}^T \cdots \gamma_{wQ}^T]^T$ and every other component defined identically as the counterpart of the model (14) where level-q covariates W_q have fixed effects γ_{wq}. The corresponding joint model (1) has $X_q = I_{p_q} \otimes [W_q^T \; W_{q+1}^T \; \cdots, W_Q^T]$ and everything else the same as previously defined.

The next section illustrates an application to three-level large-scale survey data subject to missingness at all levels. We illustrate MLE *on* Y^{mi}, which is more generally applicable than MLE *on* Y_{obs} and the hybrid method. Estimation and multiple imputation are carried out by C programs written by the first author. The imputation program uses a random number generating library of C routines, RANDLIB 1.3 by Barry W. Brown, James Lovato, Kathy Russell and John Venier. Analysis of imputed data and complete-case analysis are carried out by HLM 7 (Raudenbush, Bryk, Cheong, Congdon, and du Toit, 2011). The convergence criterion is the difference in the observed log-likelihoods between two consecutive iterations less than 10^{-5}. The statistical significance is discussed at a significance level $\alpha = 0.1$. The user-friendly two-level program that implements MLE *on* Y^{mi} is expected to be released to the public in software package HLM 7 in the year 2014. The user-friendly three-level program is under development at the time of writing this manuscript.

5 Illustrative Examples

In this section, we aim to identify the determining factors of body mass index (BMI) during childhood that may span three levels, occasions nested within a child attending a school, via analysis of the Early Childhood Longitudinal Study Kindergarten Cohort of 1998 (ECLS-K, Tourangeau et al., 2009). Specifically we consider ethnic and social disparities in the growth of BMI, and ask how environmental exposures such as television watching and school quality are associated with growth in BMI.

The ECLS-K is a nationally representative sample of 21260 kindergartners in the United States who attended 1018 schools in 1998. The study followed the children in fall-kindergarten (K) of 1998, spring-K of 1999, fall-first grade (G1) of 1999, spring-G1 of 2000, spring-third grade (G3) of 2002, spring-fifth grade (G5) of 2004 and spring-eighth grade (G8) of 2007. Due to cost constraints, a random subsample (41% to 54%) of students transferring schools were followed from K to G5. Furthermore, the fall-G1 data collection was limited to 27% of base-year students in a 30% subsample of the schools. Therefore, the ECLS-K contains many item- and unit-nonresponses. For example, only 5044 first graders had their BMI measured in fall of 1999. Consequently, researchers have analyzed the ECLS-K without the third wave in longitudinal studies of obesity (Gable et al., 2007, Bhargava, Jolliffe, and Howard, 2008, Danner, 2008). With the G8 data available since 2009, the longitudinal analysis demands challenges as less than 7% of the children attended the same school from K to G8.

A longitudinal analysis of the ECLS-K should involve all seven waves of data to yield efficient analysis. Furthermore, missing data may be present at multiple levels. The approach in this paper enables all waves and available data to be analyzed for efficient and unbiased inferences. The “all available data” include children with item- as well as unit-nonresponse because a child with time-varying characteristics missing but individual or school characteristics observed strengthens inferences at higher levels (Shin and Raudenbush, 2011, Shin, 2012). Mobile students transferring schools are nested within their original schools in fall-K and analyzed.

Following the previous studies of the ECLS-K (Datar and Sturm, 2004, Sturm and Datar, 2005, Danner, 2008, Bhargava et al., 2008), we analyze the raw BMI as a ratio of body weight in kg to height in meters squared. Table 3 summarizes the data for analysis of 21,210 children who attended 1,018 schools in 1998 after dropping 50 children with most characteristics missing including gender and race. Also dropped are 6 eighth-grade BMIs ranging 98 to 207 that are influential on the fitted regression and 13 extraneous heights and weights such as a 20-pound weight and a height reduced by more than 10 inches. After dropping the influential and extraneous observations, the standard deviation of G8 BMIs reduces from 6.29 to 5.29. With 7 occasions nested within most children, there are a total of 148,451 occasions at level 1 nested within 21,210 children at level 2 attending 1,018 schools at level 3. BMI and the daily number of hours spent watching television (TV) are time varying (Gable et al., 2007, Bhargava et al., 2008, Danner, 2008). The BMI ranges 7.1 to 57.5. To produce TV, maximum daily television viewing hours exceeding 7 per weekday and 10 per weekend day were set to 7 and 10 hours, respectively. TV was measured in spring-K for the first time and then once in every other data collection. It is unreasonable to think that the TV values between fall-K and spring-K for each child are different enough to treat all the values missing in fall-K. The analysis uses the TV measured in spring-K for the first two data collections. In addition, the analysis considers six dummy time indicators from spring-K to G8 to control for the natural growth in BMI at level 1; base-year home neighborhood safety (HOMESAFETY), base-year age in months (AGE), birth weight in pounds (BIRTHWEIGHT), base-year socioeconomic status (SES), a female indicator (FEMALE) and six race ethnicity indicators at child level or

Table 3: Data for analysis. K and Gn stand for kindergarten and nth grade respectively.

Level	Variable	Description	Mean (SD)	Missing (%)
Time	BMI	body mass index		
	Fall-K		16.27 (2.20)	2219 (10)
	Spr.-K		16.40 (2.30)	1450 (7)
	Fall-G1		16.62 (2.60)	16170 (76)
	Spr.-G1		16.90 (2.86)	5805 (27)
	Spr.-G3		18.66 (3.88)	7476 (35)
	Spr.-G5		20.57 (4.75)	10241 (48)
	Spr.-G8		22.80 (5.29)	12450 (59)
	TV	daily TV viewing in hours		
	Fall-K		2.01 (1.08)	2772 (13)
	Spr.-K		2.01 (1.08)	2772 (13)
	Fall-G1		2.44 (1.65)	16374 (77)
	Spr.-G1		1.88 (1.25)	6018 (28)
	Spr.-G3		1.91 (1.22)	8123 (38)
	Spr.-G5		2.05 (1.22)	10468 (49)
	Spr.-G8		3.09 (1.87)	12227 (58)
Child	HOMESAFETY	safety around home	1.66 (0.55)	2316 (11)
	AGE	age in months	68.41 (4.35)	2132 (10)
	BIRTHWEIGHT	birth weight in lb	6.91 (1.35)	1472 (7)
	SES	socioeconomic status	0.00 (0.80)	1088 (5)
	FEMALE	1 if female	0.49 (0.50)	0 (0)
	BLACK	1 if African-American	0.15 (0.36)	0 (0)
	HISPANIC	1 if Hispanic	0.18 (0.38)	0 (0)
	ASIAN	1 if Asian	0.06 (0.25)	0 (0)
	PACIFIC	1 if pacific islander	0.01 (0.10)	0 (0)
	ALASKAN	1 if American Indian/Alaskan	0.02 (0.13)	0 (0)
	OTHER	1 if multiracial or others	0.03 (0.16)	0 (0)
School	GRAFFITI	Graffiti around school	0.55 (0.72)	262 (26)
	PRIVATE	1 if private school	0.26 (0.44)	0 (0)

level 2; and base-year school neighborhood safety (GRAFFITI, the amount of graffiti around school) and a private school indicator (PRIVATE) at school level or level 3.

An unsafe neighborhood is associated with elevated BMI among adults (Shin and Raudenbush, 2007). HOMESAFETY has three scales: not safe or low (0); somewhat safe or medium (1); and very safe or high (2) while GRAFFITI, the lower the safer, has four scales: none (0); a little (1); some (2); and a lot (3). Preliminary analysis shows that higher-order than linear association between the safety factors and BMI is unlikely. BMI and TV miss 38 % and 44 % of their values overall, and 76% and 77% in fall-G1, respectively. HOMESAFETY, AGE, BIRTHWEIGHT and SES miss 5 to 11 % while GRAFFITI is missing for 26% of the schools. Complete-case analysis entails removing the 262 schools with missing GRAFFITI and dropping 5,381 students attending the schools and their data from analysis. The resulting inference is inefficient and may be considerably biased as will be illustrated below. The 2,166 missing birth weights in fall-K were recovered from later data collections. The 21,210 students are 49% female and 55 % white. Out of the 1018 schools, 26 % are private. The mean BMI grows with acceleration until G5.

5.1 Random Intercept Model

The analysis aims to efficiently identify the environmental factors of childhood BMI such as television watching and school quality after controlling for natural growth as well as ethnic and social disparities in BMI. At level 1, over time nested within children, we model change in BMI as a function of incompletely observed TV and time. At level 2, between children nested within schools, we include incompletely observed measure of the safety around the home, age, birth weight and socioeconomic status, and completely observed female and race ethnicity indicators. At level 3, between schools, we include an incompletely observed measure of the safety of the school by GRAFFITI and a completely observed indicator for private school. In terms of our general model (30), we therefore have $Q = 3$, $R = BMI$, $A = TV$, $Y_2^T = [HOMESAFETY\ \ AGE\ \ BIRTHWEIGHT\ \ SES]$, $Y_3 = GRAFFITI$, $D_1 = 1$, $D_2 = 1$, $D_3 = 1$, $W_1^T = [T2\ \ T3\ \ T4\ \ T5\ \ T6\ \ T7]$, $W_2^T = [FEMALE\ \ BLACK\ \ HISPANIC\ \ ASIAN\ \ PACIFIC\ \ ALASKAN\ \ OTHER]$ and $W_3^T = [1\ \ PRIVATE]$ where $T2$ through $T7$ are dummy indicators for spring-K through spring-G8. Rather than subjecting the mean growth in BMI to a polynomial curve (Bhargava et al. 2008; Danner 2008), W_1 controls for it as the difference in mean BMIs between each time point and fall-K.

Table 4 displays the output. Age and birth weight have been centered around the respective sample means. The complete-case analysis of the desired model (30) under column heading "CC" involves 12446 children attending 706 schools who have both BMI and TV observed at one or more occasions. The children have a total of 55082 occasions. The next column under "$MLE\ on\ Y^{mi}$" presents the $MLE\ on\ Y^{mi}$ that uses five imputations based on the corresponding unconstrained joint model (1) for $\pi_1 = \begin{bmatrix} \pi^1_{RR} & \pi^1_{RA} \\ \pi^1_{AR} & \pi^1_{AA} \end{bmatrix}$, $\pi_2 = \begin{bmatrix} \pi^2_{RR} & \pi^2_{RA} & \pi^2_{R2} \\ \pi^2_{AR} & \pi^2_{AA} & \pi^2_{A2} \\ \pi^2_{2R} & \pi^2_{2A} & \pi^2_{22} \end{bmatrix}$ and $\pi_3 = \begin{bmatrix} \pi^3_{RR} & \pi^3_{RA} & \pi^3_{R2} & \pi^3_{R3} \\ \pi^3_{AR} & \pi^3_{AA} & \pi^3_{A2} & \pi^3_{A3} \\ \pi^3_{2R} & \pi^3_{2A} & \pi^3_{22} & \pi^3_{23} \\ \pi^3_{3R} & \pi^3_{3A} & \pi^3_{32} & \pi^3_{33} \end{bmatrix} = \begin{bmatrix} \pi^3_{RR} & \pi^3_{RC} \\ \pi^3_{CR} & \pi^3_{CC} \end{bmatrix}$ of respective dimensions 2-by-2, 6-by-6 and 7-by-7. The number of covariance components between R and C is $\sum_{q=1}^{3} p_{Dq} \sum_{r=1}^{q} p_r - p_D = 12$ while the desired hierarchical model has $p - 1 = 6$ effects of C on R. Consequently, 6 covariance components of the unconstrained joint model are extraneous for subsequent analysis. An asterisk '*' marks statistical significance at a significance level $\alpha = 0.1$. The CC standard errors are 8 to 83% higher than those under the $MLE\ on\ Y^{mi}$. Under the CC, females have 0.09 units higher than do males, and pacific islanders and students of multiple or other races have 1.16 and 0.27 units higher than do white counterparts, respectively, in BMI on average while these effects are insignificant under the $MLE\ on\ Y^{mi}$, controlling for other covariates. The effect estimates for females and pacific islanders under the CC are 25 and 5 times higher than their counterparts under the $MLE\ on\ Y^{mi}$. The effect estimates for TV, SES, BLACK, ALASKAN and OTHER under the CC are also 10 to 67 % higher than their $MLE\ on\ Y^{mi}$ counterparts. The CC variances are comparatively underestimated.

Table 4: Random-intercept model (30) by complete-case analysis (CC), by the $MLE\ on\ Y^{mi}$ and by the $MLE\ on\ Y_{obs}$ based on the unconstrained joint model (1), and random-coefficients model (30) by the $MLE\ on\ Y^{mi}$. Statistical significance marked by '*' at a significance level $\alpha = 0.1$.

	Estimate (Std. Err.)			
	CC	$MLE\ on\ Y^{mi}$	$MLE\ on\ Y_{obs}$	$MLE\ on\ Y^{mi}$
Covariate			Unconstrained	Random Slope
Intercept	15.88 (0.11)*	15.94 (0.08)*	11.59 (0.37)*	15.93 (0.09)*
TV	0.10 (0.01)*	0.08 (0.01)*	0.18 (0.01)*	0.09 (0.01)*
T2	0.13 (0.02)*	0.13 (0.02)*	0.13 (0.02)*	0.13 (0.02)*
T3	0.34 (0.04)*	0.32 (0.03)*	0.28 (0.03)*	0.33 (0.03)*
T4	0.63 (0.03)*	0.63 (0.02)*	0.64 (0.02)*	0.63 (0.02)*
T5	2.36 (0.03)*	2.36 (0.02)*	2.37 (0.02)*	2.36 (0.03)*
T6	4.21 (0.03)*	4.24 (0.02)*	4.23 (0.02)*	4.23 (0.02)*
T7	6.38 (0.03)*	6.40 (0.02)*	6.29 (0.03)*	6.39 (0.03)*
HOMESAFETY	-0.01 (0.05)	0.00 (0.04)	0.01 (0.04)	-0.01 (0.04)
AGE	0.03 (0.01)*	0.03 (0.00)*	0.03 (0.00)*	0.03 (0.00)*
BIRTHWEIGHT	0.31 (0.02)*	0.32 (0.01)*	0.32 (0.02)*	0.32 (0.01)*
SES	-0.30 (0.04)*	-0.27 (0.03)*	-0.28 (0.03)*	-0.27 (0.03)*
FEMALE	0.09 (0.05)*	0.00 (0.04)	0.02 (0.04)	0.01 (0.04)
BLACK	0.52 (0.09)*	0.47 (0.06)*	0.35 (0.06)*	0.48 (0.07)*
HISPANIC	0.61 (0.08)*	0.62 (0.06)*	0.55 (0.06)*	0.62 (0.06)*
ASIAN	0.06 (0.13)	-0.07 (0.09)	-0.09 (0.09)	-0.10 (0.09)
PACIFIC	1.16 (0.37)*	0.25 (0.20)	0.16 (0.20)	0.31 (0.20)
ALASKAN	0.66 (0.20)*	0.53 (0.17)*	0.47 (0.16)*	0.53 (0.17)*
OTHER	0.27 (0.16)*	0.16 (0.12)	0.15 (0.13)	0.16 (0.13)
GRAFFITI	0.04 (0.05)	0.00 (0.04)	0.00 (0.04)	0.00 (0.04)
PRIVATE	-0.03 (0.07)	-0.07 (0.06)	-0.04 (0.06)	-0.07 (0.06)
τ_3	0.15 (0.03)	0.17 (0.02)	0.20 (0.03)	$\begin{bmatrix} 0.18(0.03) & -0.07(0.05) \\ -0.07(0.05) & 0.18(0.12) \end{bmatrix}$
τ_2	6.95 (0.10)	7.01 (0.08)	6.98 (0.08)	6.98 (0.07)
τ_1	3.08 (0.02)	3.07 (0.02)	3.05 (0.02)	3.07 (0.02)

The estimated natural growths of BMI under CC and $MLE\ on\ Y^{mi}$ are close to each other. From the $MLE\ on\ Y^{mi}$, a white student having mean age, birth weight and socioeconomic status who does not watch TV has 15.94 BMI units on average in fall-K. The 6-month BMI growth is 0.13 units in spring-K and then accelerates to 0.19 units (0.32-0.13) in fall-G1, to 0.31 units (0.63-0.32) in spring-G1, to 0.43 units [(2.36-0.63)/4] until spring-G3 and to 0.47 units [(4.24-2.36)/4] until spring-G5, and then decelerates to 0.36 units [(6.40-4.24)/6] until spring-G8. A polynomial curve may not reveal the details in growth. Controlling for the natural growth, and demographic individual and organizational school characteristics, a one-hour increment in daily TV viewing elevates child BMI by 0.08 units on average, 64% of the 6-month BMI growth in kindergarten. Each month in base-year age and an additional pound in birth weight contribute to 0.03 and 0.32 unit increases in BMI, respectively, while one unit increase in SES lowers the child BMI by 0.27 units on average, ceteris paribus. Black, Hispanic, and American Indian or Alaskan students have 0.47, 0.62 and 0.53 units higher, respectively, than do white counterparts in BMI on average, controlling for other covariates.

The next column under "MLE *on* Y_{obs} Unconstrained" illustrates how biased the resulting inferences may be relative to those of the desired model (30) under the MLE *on* Y^{mi} if we are to directly transform the corresponding unconstrained joint model (1) to the desired model. The transformed parameters are

$$\begin{bmatrix} \sum_{q=1}^{3} \pi_{AA}^{q} & \pi_{A2}^{2} + \pi_{A2}^{3} & \pi_{A3}^{3} \\ \pi_{2A}^{2} + \pi_{2A}^{3} & \pi_{22}^{2} + \pi_{22}^{3} & \pi_{23}^{3} \\ \pi_{3A}^{3} & \pi_{32}^{3} & \pi_{33}^{3} \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{bmatrix} = \begin{bmatrix} \sum_{q=1}^{3} \pi_{AR}^{q} \\ \pi_{2R}^{2} + \pi_{2R}^{3} \\ \pi_{3R}^{3} \end{bmatrix}, \tag{31}$$

$\tau_1 = \pi_{RR}^{1} - \gamma_1^2 \pi_{AA}^{1}$, $\tau_2 = \pi_{RR}^{2} - [\gamma_1 \ \gamma_2^T] \begin{bmatrix} \pi_{AA}^{2} & \pi_{A2}^{2} \\ \pi_{2A}^{2} & \pi_{22}^{2} \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix}$ and $\tau_3 = \pi_{RR}^{3} - \gamma^T \pi_{CC}^{3} \gamma$. We compare the estimates to the counterparts under the MLE *on* Y^{mi}. Most strikingly, the key environmental effect of television watching (TV) under investigation increases by 126% to 0.18. The gaps in mean BMIs of black, Hispanic and American Indian or Alaskan students relative to white students are noticeably underestimated, and so are the intercept and the effects of T3 and T7. The level-1 and -2 error variances are understated while the level-3 error variance is over-represented. This example is comparatively benign with only one covariate, TV, at level 1 and four covariates at level 2 subject missingness. With more covariates subject to missingness at nested levels, this method has the potential to produce severely biased inferences. To correctly apply the MLE *on* Y_{obs} for the desired hierarchical model (30), the transformation should follow estimation of the joint model under constraints (29): $\pi_{AR}^{1} = \pi_{AA}^{1} \gamma_1$, $\begin{bmatrix} \pi_{AR}^{2} \\ \pi_{2R}^{2} \end{bmatrix} = \begin{bmatrix} \pi_{AA}^{2} & \pi_{A2}^{2} \\ \pi_{2A}^{2} & \pi_{22}^{2} \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix}$ and $\pi_{CR}^{3} = \pi_{CC}^{3} \gamma$ according to Corollary 4.3. To see how the constraints work, replace the right hand side in Equation (31) with the constraints.

5.2 Random-Coefficients Model with Missing Data

The analysis above reveals that black, Hispanic, and American Indian or Alaskan students have elevated BMIs relative to white counterparts on average controlling for other covariates in the model. The minority students may attend lower-quality schools than those that white counterparts attend which, by hypothesis, contributes to the disparity in BMI. School quality may be indicated by school characteristics such as school safety, school-mean socioeconomic status, contents of school meals, physical education time and school sector (Datar and Sturm 2004; Gable et al. 2007; Bhargava et al. 2008). If this hypothesis is true, then the minority students may have a randomly varying effect on BMI across schools of different qualities. Among the minority students, Hispanic students stand out in BMI. Overall, Hispanic students are half as likely to attend private schools as white students. They also attend schools having about three times as much graffiti around as those that white students attend on average.

The random-intercept model above is extended to a random-coefficients model where the Hispanic indicator has a random effect on BMI across schools. The desired model (30) has $D_3 = \begin{bmatrix} 1 \\ HISPANIC \end{bmatrix}$, $\tau_3 = \begin{bmatrix} \tau_{300} & \tau_{301} \\ \tau_{301} & \tau_{311} \end{bmatrix}$ and all other components identical to those of the random intercept model. The corresponding unconstrained joint model (1) has an 8-by-8 covariance matrix $\pi_3 = \begin{bmatrix} \pi^3_{R0R0} & \pi^3_{R0R1} & \pi^3_{R0C} \\ \pi^3_{R1R0} & \pi^3_{R1R1} & \pi^3_{R1C} \\ \pi^3_{CR0} & \pi^3_{CR1} & \pi^3_{CC} \end{bmatrix}$ and all other parameters identical to those of the joint model corresponding to the random-intercept model. The complete-case analysis produced virtually identical estimates as those under the CC in Table 4. This is not surprising in that the slope of HISPANIC does not vary significantly across schools (variance estimate=0.23, standard error=0.17, p-value=0.24). The $MLE\ on\ Y^{mi}$ based on $m = 5$ imputations yields the estimates under "$MLE\ on\ Y^{mi}$ Random Slope" in the Table. Except for the level-3 covariance matrix, the estimates are close to the counterparts under the $MLE\ on\ Y^{mi}$. The slope for the Hispanic indicator seems to vary at most modestly across schools (slope= 0.18, standard error= 0.12). To test the null hypothesis that $\tau_{311} = 0$ which implies $\tau_{301} = 0$, let θ_F and θ_R be the parameters of the random-coefficients (full) and -intercept (reduced) models, and $\hat{\theta}^t_F$ and $\hat{\theta}^t_R$ be the ML estimates, respectively, given the tth imputation for $t = 1, \cdots, m$. The log likelihoods $l(\hat{\theta}^t_F)$ and $l(\hat{\theta}^t_R)$ evaluated at the ML $\hat{\theta}^t_F$ and $\hat{\theta}^t_R$, respectively, given the tth imputation yield $d_t = 2[l(\hat{\theta}^t_F) - l(\hat{\theta}^t_R)]$. The test statistic recommended by Li, Meng, Raghunathan, and Rubin (1991a) is $D = \left[\bar{d}/2 - (m+1)(m-1)^{-1}r\right]/(1+r) = 0.53$ where $\bar{d} = \sum_{t=1}^m d_t/m$ and $r = (1 + m^{-1})\left[\sum_{t=1}^m \left(\sqrt{d_t} - \overline{\sqrt{d}}\right)^2/(m-1)\right]$ for $\overline{\sqrt{d}} = \sum_{t=1}^m \sqrt{d_t}/m$ (Schafer, 1997). The p-value is $P(F_{2,\nu} > D) = 0.59$ where $F_{2,\nu}$ is a random variable from the F distribution with 2 numerator and ν denominator degrees of freedom for $\nu = (m-1)(1+1/r)^2/2^{3/m} = 974$. This test yields an approximate range of p-values between twice and one half the computed value (Li et al., 1991a, Schafer, 1997). The computed p-value 0.59 gives enough precision to conclude the random intercept (null) model. Therefore, we do not find evidence that the slope for the Hispanic indicator varies randomly across schools.

The unconstrained joint model identifies 18 covariance components between R and C while the desired random coefficient model has 6 effects of C on R. Consequently, 12 covariance components between R and C are extraneous for subsequent analysis. Constraints to identify the desired model via $MLE\ on\ Y_{obs}$ are $\pi^1_{AR} = \pi^1_{AA}\gamma_1$, $\begin{bmatrix} \pi^2_{AR} \\ \pi^2_{2R} \end{bmatrix} = \begin{bmatrix} \pi^2_{AA} & \pi^2_{A2} \\ \pi^2_{2A} & \pi^2_{22} \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix}$, $\pi^3_{CR0} = \pi^3_{CC}\gamma$ and $\pi^3_{R1C} = 0$ by Corollary 4.3.

6 Discussion

This paper presented methods for efficient and unbiased analysis of a Q-level hierarchical general linear model given incomplete data with a general missing pattern at any of the Q levels. Our general approach uniformly expresses the Q-level model for $Q \geq 1$ that greatly facilitates extension of existing single-level and two-level efficient missing data methods to general Q-level data; reexpresses the desired model as a joint distribution of the variables, including the outcome, that are subject to missingness conditional on all of the covariates that are completely observed; and efficiently estimates the joint distribution. This approach confronts two major challenges. As the number of Q levels, the number, p, of variables subject to missingness and the number, p_D, of random coefficients increase in the hierarchical model, the joint distribution may become extremely high dimensional and difficult to estimate well. Moreover, the joint model, in general, over-identifies the desired hierarchical model. The problem of over-identification can grow severe as levels, covariates, and random coefficients are added to the hierarchical model. The consequence is that the overidentified hierarchical model may produce considerably biased inferences as was illustrated in this paper. To overcome the computational challenges, we derived, within each iteration of the EM algorithm, recursive Q-step computation formulas for efficient estimation of the joint distribution where computation at each step involves single-level data only given higher-level computation components. The consequence is efficient computation that is not excessively burdened with regard to Q, p and p_D. Furthermore, we showed how to impose constraints on the joint distribution within the framework of the Q-level hierarchical model in a way that is uniform without regard to Q; and in a way that produces unbiased and efficient analysis of the hierarchical model.

This paper considered three methods for efficient handling of missing data: $MLE\ on\ Y_{obs}$, $MLE\ on\ Y^{mi}$ and the hybrid method. Given a Q-level hierarchical model with incomplete data, $MLE\ on\ Y_{obs}$ constrains the joint model to be a one-to-one transformation of the desired hierarchical model for unbiased analysis; efficiently estimates the constrained joint model by ML; and transforms it to the hierarchical model. Consequently, $MLE\ on\ Y_{obs}$ is tailored to estimation of the desired hierarchical model. On the other hand, $MLE\ on\ Y^{mi}$ generates multiple imputation given the unconstrained joint model estimated by ML, allowing the user to impose the desired constraints when using conventional software to analyze the imputed data. Therefore, $MLE\ on\ Y^{mi}$ is more generally applicable by enabling the data analyst to explore, in addition, contextual effects and interaction effects involving covariates having random coefficients. Theorem 4.1 provides the scope of hierarchical models that can be explored under the joint model. When $MLE\ on\ Y^{mi}$ is desired, but produces an unconstrained joint model that is extremely high dimensional and thus difficult to estimate well, the hybrid method of imputation following estimation of a partially constrained joint model enables estimation of fewer parameters and thus reduces computational burden in estimation. The Corollaries to Theorem 4.1 provide the scope of hierarchical models that may be explored by the data analyst under the partially constrained joint model.

We have illustrated the $MLE\ on\ Y^{mi}$ by a longitudinal analysis of ECLS-K for $Q = 3$ and compared it to the complete-case analysis that was shown to be relatively inefficient and subject to biased inferences. We have also compared the $MLE\ on\ Y^{mi}$ to the $MLE\ on\ Y_{obs}$ based on the unconstrained joint model that revealed the potential to produce substantially biased inferences. The proliferation of extraneous parameters was comparatively moderate with $Q = 3$, $p = 7$ and $p_D = 3$. We have imposed six constraints to generate the results via the $MLE\ on\ Y^{mi}$ in Table 4. For a model having larger Q, p or p_D, it may be desirable to use the hybrid method that reduces the computational burden of $MLE\ on\ Y^{mi}$ and, at the same time, broadens the scope of $MLE\ on\ Y_{obs}$ by allowing an analyst to explore multiple hierarchical models for the correct specification of the desired model. We may also take advantage of extra variables not of direct interest in the desired hierarchical model, but highly correlated with variables subject to missingness to more precisely impute missing data at multiple levels (Shin and Raudenbush, 2007).

The ECLS-K has a great majority of students transferring schools. Thus, it may be more appropriate to consider a cross-classified model for the analysis that relaxes the strict hierarchy of nesting a student within a single school. Estimation of a cross-classified model is challenging because the growths in the outcome of students while attending the same schools become dependent to produce a complicated network of dependence among children and schools (Raudenbush and Bryk, 2002). All schools for each child and all children for each school may have to be analyzed at once to fully account for the dependence. With many covariates subject to missingness at multiple levels, the joint model of variables subject to missingness may be too highly dimensioned to estimate well. Therefore, a valuable future research topic is development of a method for efficient estimation of a cross-classified model given incomplete data.

One restriction of the general Q-level hierarchical model is that the covariates having random effects should be completely observed. If the covariates are subject to missingness, they should appear on the left hand side of the corresponding joint model for efficient handling of the missing data as well as on the right hand side of the model for estimation of the random coefficients. Such a joint model is not multivariate normal, and the factorization under joint normality that leads to the desired conditional hierarchical linear model does not apply. Consequently, the ML approach is challenging. Relaxing this assumption is beyond the scope of the current paper.

It took 21 seconds to complete each iteration in estimating each joint model in Table 4 on a 2.8 GHz laptop computer that has 8 GB memory. The estimated random intercept and coefficients models took more than 5 hours to converge at 867th and 894th iterations, respectively. No attempt to accelerate the convergence has been made. The convergence criterion is the difference in the observed log-likelihoods between two consecutive iterations less than 10^{-5}. In some of our two-level test runs, we compared computation times for estimation of a joint model between our program with a convergence criterion of the difference in log-likelihoods between two consecutive iterations less than 10^{-6} (Shin and Raudenbush, 2007), and an alternative program with a convergence criterion of the percentage difference in log-likelihoods between two consecutive iterations less than 10^{-6} and the Aitken acceleration (Aitken, 1926), the alternative program converged not only to practically identical estimates and standard errors, but up to 90% faster than did our program in terms of the number of iterations. Considerable saving in computation time is anticipated with the likewise acceleration in the three-level applications. It took us about 2 minutes to generate a single imputation for the results in Table 4. So far, we have developed a three-level program implementing the $MLE\ on\ Y^{mi}$ only and thus cannot compare the computation times between $MLE\ on\ Y^{mi}$ and $MLE\ on\ Y_{obs}$.

In Section 5.2, we found no evidence that the slope for the Hispanic student indicator varies randomly across schools, based on the test statistic recommended by Li et al. (1991a). This test provides an approximate range of p-values between twice and one half the computed p-value. More accurate p-values may be obtained at the expense of extra computational effort (Li, Raghunathan, and Rubin, 1991b, Meng and Rubin, 1992, Schafer, 1997, Little and Rubin, 2002). Because the corollaries to Theorem 4.1 establish one-to-one correspondence between hierarchical model (30) and joint model (1), the $MLE\ on\ Y_{obs}$ enables an alternative likelihood ratio test between the two analyst's models directly based on their constrained joint models.

Our illustrative examples in section 5 are based on a large sample. The performance of our estimators in terms of bias and efficiency involving a small sample is yet to be assessed. Therefore, simulation studies on the small-sample performance of our methods will be a useful future research area.

The analysis in this paper involved discrete covariates, the safety factors, subject to missingness at levels 2 and 3. Although it is improper for the normal linear joint model to describe the marginal distribution for the discrete factors, the implied conditional distribution is the desired hierarchical model. An advantage is that it allows the discrete covariates subject to missingness to be analyzed by the efficient missing data method (Schafer, 1997, Shin and Raudenbush, 2007). In addition, the impact of the joint distribution assumptions on the desired conditional model by the $MLE\ on\ Y^{mi}$ is comparatively weak because the distributional assumptions do not affect the observed data. A valuable future extension of this approach is to a hierarchical generalized linear model given incomplete data.

Appendix

Proof of Proposition 3.1. The model (8) implies

$$\begin{aligned}
A^q_{mobs} &= \left[\begin{array}{l} O^T_{qm}\psi^{-1}_{qm}(d_{qmobs} - Z_{-qqmobs}\tilde{\epsilon}_{-qqm}) \\ \Omega^{-1}_{qm}(\tilde{\epsilon}_{-qqm} - \tilde{\tilde{\epsilon}}_{-qqm}) \\ A^{q+1}_{mobs} - B^{q+1}_{mobs}\Phi^{q(q+1)T}_m\Omega^{-1}_{q11m}(\tilde{\epsilon}_{-qqm} - \tilde{\tilde{\epsilon}}_{-qqm}) \end{array}\right], \qquad (32)\\
B^q_{mobs} &= \left[\begin{array}{ccc} O^T_{qm}V^{-q11}_{mobs}O_{qm} & O^T_{qm}V^{-q11}_{mobs}Z_{-qqmobs} & O^T_{qm}V^{-q12}_{mobs}Z^{q+1}_{mobs} \\ & Z^T_{-qqmobs}V^{-q11}_{mobs}Z_{-qqmobs} & Z^T_{-qqmobs}V^{-q12}_{mobs}Z^{q+1}_{mobs} \\ & & Z^{(q+1)T}_{mobs}V^{-q22}_{mobs}Z^{q+1}_{mobs} \end{array}\right] \qquad (33)
\end{aligned}$$

for a symmetric matrix (33) showing the upper triangular components only where Ω_{qm}, $\tilde{\epsilon}_{-qqm}$, $\tilde{\tilde{\epsilon}}_{-qqm}$ and V^{-q11}_{mobs} in Equations (12), and

$$\begin{aligned}
V^{-q12}_{mobs}Z^{q+1}_{mobs} &= -\psi^{-1}_{qm}Z_{-qqmobs}\Delta^{-1}_{qm}\Omega^{-1}_{qm}\Phi^{q(q+1)}_m B^{q+1}_{mobs}, \\
Z^{(q+1)T}_{mobs}V^{-q22}_{mobs}Z^{q+1}_{mobs} &= B^{q+1}_{mobs} + B^{q+1}_{mobs}\Phi^{q(q+1)T}_m(\Omega^{-1}_{qm} - \Omega^{-1}_{qm}\Delta^{-1}_{qm}\Omega^{-1}_{qm})\Phi^{q(q+1)}_m B^{q+1}_{mobs}
\end{aligned}$$

depend on level-q $Z_{-qqmobs}$ and d_{qmobs} only, given A^{q+1}_{mobs}, B^{q+1}_{mobs} and θ. ∎

Proof of Theorem 3.2. The initial step is to compute A^Q_{mobs} and B^Q_{mobs} in Equations (11). Suppose that $\tilde{b}^q_m$ and $\tilde{\Pi}^{qq}_m$ may be computed involving level-q data only given A^{q+1}_{mobs}, B^{q+1}_{mobs} and θ. Then, $\tilde{b}^{q-1}_m$ and $\tilde{\Pi}^{(q-1)(q-1)}_m$ may be computed based on level-$(q-1)$ data only given A^q_{mobs}, B^q_{mobs} and θ by Proposition 3.1. ∎

Proof of Proposition 3.3. Equation (8) implies

$$\begin{aligned}
F^q_{mobs} &= \left[\begin{array}{l} X^T_{qmobs}\psi^{-1}_{qm}(d_{qmobs} - Z_{-qqmobs}\tilde{\epsilon}_{-qqm}) \\ F^{q+1}_{mobs} - H^{(q+1)T}_{mobs}\Phi^{q(q+1)T}_m\Omega^{-1}_{qm}(\tilde{\epsilon}_{-qqm} - \tilde{\tilde{\epsilon}}_{-qqm}) \end{array}\right], \qquad (34)\\
G^q_{mobs} &= \left[\begin{array}{cc} X^T_{qmobs}V^{-q11}_{mobs}X_{qmobs} & X^T_{qmobs}V^{-q12}_{mobs}X^{q+1}_{mobs} \\ (X^T_{qmobs}V^{-q12}_{mobs}X^{q+1}_{mobs})^T & X^{(q+1)T}_{mobs}V^{-q22}_{mobs}X^{q+1}_{mobs} \end{array}\right] \qquad (35)
\end{aligned}$$

$$H^q_{mobs} = \left[\begin{array}{cc} Z^T_{qmobs} V^{-q11}_{mobs} X_{qmobs} & Z^T_{qmobs} V^{-q12}_{mobs} X^{q+1}_{mobs} \\ (X^T_{qmobs} V^{-q12}_{mobs} Z^{q+1}_{mobs})^T & Z^{(q+1)T}_{mobs} V^{-q22}_{mobs} X^{q+1}_{mobs} \end{array} \right] \tag{36}$$

where Ω_{qm}, $\tilde{\epsilon}_{-qqm}$, $\tilde{\tilde{\epsilon}}_{-qqm}$ and V^{-q11}_{mobs} in Equations (12), $V^{-q12}_{mobs} Z^{q+1}_{mobs}$ in Equation (33),

$$\begin{array}{lcl} V^{-q12}_{mobs} X^{q+1}_{mobs} & = & -\psi^{-1}_{qm} Z_{-qqmobs} \Delta^{-1}_{qm} \Omega^{-1}_{qm} \Phi^{q(q+1)}_m H^{q+1}_{mobs}, \\ X^{(q+1)T}_{mobs} V^{-q22}_{mobs} X^{q+1}_{mobs} & = & G^{q+1}_{mobs} + H^{(q+1)T}_{mobs} \Phi^{q(q+1)T}_m (\Omega^{-1}_{qm} - \Omega^{-1}_{qm} \Delta^{-1}_{qm} \Omega^{-1}_{qm}) \Phi^{q(q+1)}_m H^{q+1}_{mobs} \end{array}$$

depend on level-q $Z_{-qqmobs}$, Y_{qmobs} and X_{qmobs} only, given A^{q+1}_{mobs}, B^{q+1}_{mobs}, F^{q+1}_{mobs}, G^{q+1}_{mobs}, H^{q+1}_{mobs} and θ. $\blacksquare$

Proof of Theorem 3.4. The initial step is to compute $A^Q_{mobs} = O^T_{Qm} \pi^{-Q}_{QQm} d_{Qmobs}$, $B^Q_{mobs} = O^T_{Qm} \pi^{-Q}_{QQm} O_{Qm}$, $F^Q_{mobs} = X^T_{Qmobs} \pi^{-Q}_{QQm} d_{Qmobs}$, $G^Q_{mobs} = X^T_{Qmobs} \pi^{-Q}_{QQm} X_{Qmobs}$ and $H^Q_{mobs} = O^T_{Qm} \pi^{-Q}_{QQm} X_{Qmobs}$ for $X_{Qmobs} = X^Q_{mobs}$. Suppose that F^q_{mobs} and G^q_{mobs} can be computed from level-q data only, given A^{q+1}_{mobs}, B^{q+1}_{mobs}, F^{q+1}_{mobs}, G^{q+1}_{mobs}, H^{q+1}_{mobs} and θ. Then, F^{q-1}_{mobs} and G^{q-1}_{mobs} can be computed from level-$(q-1)$ data only, given A^q_{mobs}, B^q_{mobs}, F^q_{mobs}, G^q_{mobs}, H^q_{mobs} and θ by Proposition 3.3. $\blacksquare$

Proof of Theorem 4.1. Table 2 reveals $\sum_{q=1}^Q p_{Dq} \sum_{r=1}^q p_r - p_D$ covariance components in $(\pi^1_{RC}, \pi^2_{RC}, \cdots, \pi^Q_{RC})$ between R and C and implies $E(\epsilon_{qR}|\epsilon_{qC}) = \Gamma_q \epsilon_{qC}$ and $var(\epsilon_{qR}|\epsilon_{qC}) = \pi^q_{R|C}$ for all q to identify model (26). Conversely, Equation (26) implies $E(R|D) = 0$, $var(R|D) = \sum_{q=1}^Q D^T_q \pi^q_{RR} D_q$ and $cov(R, C|D) = \left[\sum_{q=1}^Q D^T_q \pi^q_{RA} \; \sum_{q=2}^Q D^T_q \pi^q_{R2} \cdots D^T_Q \pi^Q_{RQ} \right]$ which, along with the marginal C, is one-to-one with the joint model (1). $\blacksquare$

Fisher Information
We express $V^1_{mobs} = \sum_{r=1}^Q \sum_{i=1}^{N_{rm}} \left(\bigoplus_{q=1}^r A_{rqimobs} \times \pi_r \times \bigoplus_{q=1}^r A^T_{rqimobs} \right)$ to compute $var(\hat{\theta})$ where $A_{rqimobs} = [0 \; \cdots \; 0 \; Z^T_{rqimobs} \; 0 \; \cdots \; 0]^T$ places $Z_{rqimobs}$ for the ith level-q unit within cluster m and zeroes elsewhere for $r < Q$ and $A_{Qq1mobs} = Z_{Qqmobs}$ for $r = Q$. Let φ_q be a vector of distinct elements in π_q so that $\varphi = [\varphi^T_Q \; \varphi^T_{Q-1} \; \cdots \; \varphi^T_1]^T$ and $E_r = dvec\pi_r / d\varphi^T$. The Fisher information is

$$diag \left\{ \frac{1}{2} \sum_{m=1}^{N_Q} \left(\frac{\partial vec V^1_{mobs}}{\partial \varphi^T} \right)^T (V^{-1}_{mobs} \otimes V^{-1}_{mobs}) \frac{\partial vec V^1_{mobs}}{\partial \varphi^T}, \sum_{m=1}^{N_Q} X^{1T}_{mobs} V^{-1}_{mobs} X^1_{mobs} \right\} \tag{37}$$

for $\frac{\partial vec V^1_{mobs}}{\partial \varphi^T} = \sum_{r=1}^Q \sum_{i=1}^{N_{rm}} \left(\bigoplus_{q=1}^r A_{rqimobs} \otimes \bigoplus_{q=1}^r A_{rqimobs} \right) E_r$ and $(\frac{\partial vec V^1_{mobs}}{\partial \varphi^T})^T (V^{-1}_{mobs} \otimes V^{-1}_{mobs}) \times \frac{\partial vec V^1_{mobs}}{\partial \varphi^T} = \sum_{r'=1}^Q \sum_{i'=1}^{N_{rm}} \sum_{r=1}^Q \sum_{i=1}^{N_{rm}} E^T_{r'} \left(\bigoplus_{q'=1}^{r'} A^T_{r'q'i'mobs} V^{-1}_{mobs} \bigoplus_{q=1}^r A_{rqimobs} \right) \otimes \left(\bigoplus_{q'=1}^{r'} A^T_{r'q'i'mobs} V^{-1}_{mobs} \bigoplus_{q=1}^r A_{rqimobs} \right) E_r$.

Likelihood
The observed log-likelihood $l(\theta|d_{obs}) \propto -\frac{1}{2} \sum_{m=1}^{N_Q} (log|V^1_{mobs}| + d^{1T}_{mobs} V^{-1}_{mobs} d^1_{mobs})$ measures con-

vergence to ML for $d_{obs} = (d_{1obs}, d_{2obs}, \cdots, d_{N_Q obs})$. Recursive components are

$$
\begin{aligned}
log|V^q_{mobs}| &= \sum_{i=1}^{N_{qm}} log|\pi^q_{qqim}| + log|\Delta_{qm}| + log|\Omega_{qm}| + log|V^{q+1}_{mobs}| \qquad (38) \\
d^{qT}_{mobs} V^{-q}_{mobs} d^q_{mobs} &= d^{(q+1)T}_{mobs} V^{-(q+1)}_{mobs} d^{q+1}_{mobs} + [d_{qmobs} - Z_{-qqmobs} E(\epsilon_{-qqm}|Y^{q+1}_{mobs})] \qquad (39) \\
&\quad \times \bigoplus_{i=1}^{N_{qm}} \pi^{-q}_{qqim}(d_{qmobs} - Z_{-qqmobs}\tilde{\epsilon}_{-qqm}).
\end{aligned}
$$

References

Aitken, A. (1926): "On bernoulli's numerical solution of algebraic equations," *Proceedings of the Royal Society of Edinburgh*, 46, 289–305.

Allison, P. D. (1987): "Estimation of linear models with incomplete data," *Sociological Methodology*, 17, 71–103.

Arbuckle, J. L. (1996): "Full information estimation in the presence of incomplete data," in G. A. Marcoulides and R. E. Schumacker, eds., *Advanced Structural Equation Modeling: Issues ande Techniques*, Mahwah, NJ: Lawrence Erlbaum.

Arbuckle, J. L. (2003): *Amos 5.0 Update to the Amos 5.0 Users Guide [Computer Software and Manual]*, Chicago: Smallwaters.

Bentler, P. M. (2007): *EQS 6 Structural Equations Program Manual*, Encino, CA: Multivariate Software.

Bhargava, A., D. Jolliffe, and L. Howard (2008): "Socio-economic, behavioural and environmental factors predicted body weights and household food insecurity scores in the early childhood longitudinal study-kindergarten," *Br. J. Nutrition*, 100, 438–444.

Bingenheimer, J. and S. Raudenbush (2004): "Statistical and substantive inferences in public health: issues in the application of multilevel models," *Annual Review of Public Health*, 25, 53–77.

Danner, F. (2008): "A national lognitudinal study of the association between hours of tv viewing and the trajectory of bmi growth among us children," *Journal of Pediatric Psychology*, 33, 1100–07.

Datar, A. and R. Sturm (2004): "Physical education in elementary school and body mass index: Evidence from the early childhood longitudinal study," *Am. J. Public Health*, 94, 1501–6.

Dempster, A., N. Laird, and D. Rubin (1977): "Maximum likelihood from incomplete data via the em algorithm," *JRSS, Series B*, 76, 1–38.

Enders, C. and J. Peugh (2004): "Using an em covariance matrix to estimate structural equation models with missing data: Choosing an adjusted sample size to improve the accuracy of inferences," *Structural Equation Modeling*, 11, 1–19.

Gable, S., Y. Chang, and J. Krull (2007): "Television watching and frequency of family meals are predictive of overweight onset and persistence in a national sample of school-aged children," *Journal of the American Dietetic Association*, 107, 53–61.

Goldstein, H. (2003): *Multilevel Statistical Models*, London: Edward Arnold.

Goldstein, H. and W. Browne (2002): "Multilevel factor analysis modelling using markov chain monte carlo estimation," in G. Marcoulides and I. Moustaki, eds., *Latent variable and latent structure models*, London: Lawrence Erlbaum.

Goldstein, H. and W. Browne (2005): "Multilevel factor analysis models for continuous and discrete data," in A. Olivares and J. J. McArdle, eds., *Contemporary Psychometrics. A Festschrift to Roderick P. McDonald*, Mahwah, NJ: Lawrence Erlbaum.

Goldstein, H., J. Carpenter, M. Kenward, and K. Levin (2009): "Multilevel models with multivariate mixed response types," *Statist. Modellng*, 9, 173–197.

Goldstein, H. and D. Kounali (2009): "Multilevel multivariate modelling of childhood growth, numbers of growth measurements and adult characteristics," *JRSS, Series A*, 172, 599–613.

Li, K. H., X. Meng, T. E. Raghunathan, and D. B. Rubin (1991a): "Significance levels from repeated p-values with multiply-imputed data." *Statistica Sinica*, 1, 65–92.

Li, K. H., T. E. Raghunathan, and D. B. Rubin (1991b): "Large-sample significance levels from multiply imputed data using moment-based statistics and an f reference distribution," *JASA*, 86, 1065–73.

Little, R. and D. Rubin (2002): *Statistical Analysis with Missing Data*, New York: Wiley.

Liu, M., J. Taylor, and T. Belin (2000): "Multiple imputation and posterior simulation for multivariate missing data in longitudinal studies," *Biometrics*, 56, 1157–63.

Magnus, J. R. and H. Neudecker (1988): *Matrix Differential Calculus with Applications in Statistics and Econometrics*, New York: Wiley.

Meng, X. (1994): "Multiple-imputation inferences with uncongenial sources of input," *Statistical Science*, 9, 538–558.

Meng, X. and D. Rubin (1992): "Performing likelihood ratio tests with multiply-imputed data sets," *Biometrika*, 79, 103–111.

Muthén, B. (1993): "Latent variable modeling of growth with missing data and multilevel data," in C. M. Cuadras and C. R. Rao, eds., *Multivariate Analysis: Future Directions 2*, Amsterdam: North Holland, 199–210.

Muthén, B., D. Kaplan, and M. Hollis (1987): "On structural equation modeling with data that are not missing completely at random," *Psychometrika*, 52, 431–462.

Muthén, L. and B. Muthén (2010): *Mplus user's guide (6th ed.)*, Los Angeles, CA: Muthén and Muthén, 6th edition.

Raudenbush, S. W. and A. S. Bryk (2002): *Hierarchical Linear Models*, Newbury Park, CA: Sage.

Raudenbush, S. W., A. S. Bryk, Y. Cheong, R. Congdon, and M. du Toit (2011): *HLM 7: Hierarchical linear and nonlinear modeling*, Lincolnwood, IL: Scientific Software International.

Rubin, D. (1976): "Inference and missing data," *Biometrika*, 63, 581–592.

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

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

Schafer, J. (1999): *NORM: Multiple imputation of incomplete multivariate data under a normal model [Computer software]*, University Park: Pennsylvania State University, Department of Statistics.

Schafer, J. (2003): "Multiple imputation in multivariate problems when imputation and analysis models differ," *Statistica Neerlandica*, 57, 19–35.

Schafer, J. and R. Yucel (2002): "Computational strategies for multivariate linear mixed-effects models with missing values," *JCGS*, 11, 437–457.

Shin, Y. (2012): "Do black children benefit more from small classes? multivariate instrumental variable estimators with ignorable missing data," *JEBS*, 37, 543–574.

Shin, Y. and S. W. Raudenbush (2007): "Just-identified versus over-identified two-level hierarchical linear models with missing data," *Biometrics*, 63, 1262–68.

Shin, Y. and S. W. Raudenbush (2010): "A latent cluster mean approach to the contextual effects model with missing data," *JEBS*, 35, 26–53.

Shin, Y. and S. W. Raudenbush (2011): "The causal effect of class size on academic achievement: Multivariate instrumental variable estimators with data missing at random," *JEBS*, 36, 154–185.

Sturm, R. and A. Datar (2005): "Body mass index in elementary school children, metropolitan area food prices and food outlet density," *Public Health*, 119, 105968.

Tourangeau, K., C. Nord, T. Lê, A. Sorongon, and M. Najarian (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*, Washington, D.C.: NCES, IES, DOE, (nces 2009-004) edition.

Yucel, R. (2008): “Multiple imputation inference for multivariate multilevel continuous data with ignorable non-response,” *Philosophical Transactions of the Royal Society A*, 366, 2389–2403.