Efficient Handling of Predictors and Outcomes Having Missing Values

Yongyun Shin

In Rutkowski, L., von Davier, M., & Rutkowski, D. (Eds.),

Handbook of international large-scale assessment:

background, technical issues, and methods of data analysis,

chapter 20, p 451-479, New York, NY: CRC Press, 2013

Hierarchical organization of schooling in all nations insures that international large-scale assessment data are multilevel where students are nested within schools and schools are nested within nations. Longitudinal follow-up of these students adds an additional level. Hierarchical or multilevel models are appropriate to analyze such data (Raudenbush and Bryk 2002; Goldstein 2003). A ubiquitous problem, however, is that explanatory as well as outcome variables may be subject to missingness at any of the levels, posing the data analyst with a challenge.

This chapter explains how to efficiently analyze a two-level hierarchical linear model given incompletely observed data where students at level 1 are nested within schools at level 2. This social setting may also apply to occasions nested within individuals, students nested within nations, and schools nested within nations. The efficient missing data method we use in this chapter aims to analyze all available data (Shin and Raudenbush 2007). The "all available data" include children with item as well as unit nonresponse as they belong to a school and a nation having observed data and thus add information to strengthen inferences at higher levels (Shin and Raudenbush 2011; Shin 2012).

Section 1 clarifies the assumptions we make about missing data for efficient analysis of multilevel incomplete data. Section 2 summarizes currently available methods for analysis of multilevel incomplete data. Section 3 introduces the missing data method we use in this chapter and explains how it efficiently estimates a hierarchical linear model given incomplete data. Section 4 illustrates efficient analysis of a hierarchical linear model given the incompletely observed US data from the Programme for International Student Assessment (PISA, OECD 2007). Section 5 illustrates an analysis strategy with plausible values given the PISA data where each missing value of the outcome variable is filled in or imputed with five plausible values, but predictors may be subject to missingness. Section 6 discusses the extensions and limitations of the efficient missing data method.

1 Assumptions about Missing Data

In this chapter, we consider analysis of incompletely observed two-level data with the most common missing data pattern in education, a general missing pattern. That is, the missing data method to be introduced in section 3 efficiently handles explanatory as well as outcome variables that are subject to missingness with any missing patterns at a single level or multiple levels. Consequently, we do not distinguish different types of missing patterns produced by item or unit nonresponse.

Nearly all educational data sets are multilevel and have missing data. Until quite recently, researchers facing multilevel incomplete data analysis have dropped cases with missing values. The complete-case analysis is more problematic in multilevel analysis than it is in single-level analysis. A missing national characteristic, for example, implies deletion of not only the nation but all nested schools and students within the nation. Such analysis lowers sample sizes at multiple levels to produce inefficient inferences. Consequently, the standard errors of parameter estimates will be larger than they should, resulting in conservative hypothesis tests and excessively wide confidence intervals. In addition, the missing data patterns may be associated with the deleted data to produce biased inferences (Little and Rubin 2002). Unbiased analysis is achieved when missing data are a random sample of the complete data, so that the missing data patterns are not associated with complete data, i.e., when data are missing completely at random (MCAR, Rubin 1976). However, data MCAR is seldom a reasonable assumption.

Missing values may be imputed or filled in by ad hoc imputation methods such as a sample-mean substitution and a regression model-based prediction. The substituted or predicted values, however, under-represent the true uncertainty in the missing values to produce underestimated standard errors of parameter estimates. Consequently, the resulting hypothesis tests will be too liberal and the confidence intervals too narrow. Missing values may also be filled in by other imputation methods such as a last-observation-carry-forward method (Krueger 1999) and hot deck imputation (Little and Rubin 2002) where missing values are replaced with observed values of similar units. These single imputation methods take each imputed value as if it is the true value for subsequent complete-data analysis. The estimation, however, does not take into account uncertainty due to missing data to yield understated standard errors of parameter estimates. In general, these ad hoc imputation methods are not recommended unless missing data consist of a small fraction of complete data.

In this chapter, we shall employ two comparatively mild assumptions in many applications that data are missing at random (MAR) and that the parameters, θ, of the desired hierarchical model are distinct from the nuisance parameters, ϕ, of the missing data generating mechanism or the model for missing patterns (Rubin 1976). The MAR assumption means that missing data patterns are conditionally independent of missing data given observed data. That is, the association between missing data patterns and complete data is explained by observed data. When variables subject to missingness are highly correlated, for example, the observed data are likely to explain the association between missing data patterns and complete data to make the MAR assumption plausible (Shin and Raudenbush 2011; Shin 2012). The MAR assumption requires that we analyze all observed data for efficient analysis. The distinct parameter assumption is reasonable if there is little reason to believe that knowing the nuisance parameters ϕ provides extra information on the desired parameters θ (Schafer 1997). In regression analysis for the effect of socioeconomic status on a math achievement outcome, for example, a student may not take the exam because she is sick or because she moves to a different school due to relocation of her family. It is not reasonable to believe that knowing such a mechanism would provide more information about the desired effect. In that case, the distinct parameter assumption is reasonable. On the other hand, if low performers are more likely to miss the exam than high performers such that knowing ϕ of the missing data generating mechanism is informative about the desired effect, then the distinct parameter assumption is not reasonable. Data missing under these two assumptions are called ignorable (Little and Rubin 2002). The ignorable missing data assumption is much weaker than the MCAR assumption (Rubin 1976; Schafer 1997; Little and Rubin 2002). Note that the MCAR implies the distinct parameter assumption.

Missing data neither MAR nor MCAR are said to be not missing at random

(NMAR, Rubin 1976; Little and Rubin 2002). Under this assumption, missing patterns are associated with observed as well as missing data. A longitudinal study, for example, produces informative dropouts where the dropout patterns are associated with unobserved as well as observed outcomes (Diggle and Kenward 1994; Little 1995; Muthén et al. 2011). Consequently, both θ and ϕ have to be estimated from the joint distribution of complete data and missing patterns. This amounts to estimating, in addition to the desired hierarchical model, the model for missing patterns. Because the joint model involves missing data, the model assumptions yield parameters that are not uniquely identifiable, or parameter estimates that are not supported by observed data (Little 2009). Such a parameter may be constrained for identification or assumed to take a value in estimation. In general, little evidence exists in observed data to support such a parameter estimate. Consequently, sensitivity analysis should follow estimation of the joint distribution over the range of plausible values of the parameter (Little 1995, 2009). Therefore, analysis given data NMAR is more challenging than that given data MAR or MCAR.

In this chapter, we employ the ignorable missing data assumption that is quite plausible in many applications (Schafer 1997; Little and Rubin 2002). It is also the weakest condition under which we produce valid inferences by analyzing the desired hierarchical linear model only, i.e., by ignoring the missing data generating mechanism (Rubin 1976). The next section reviews currently available methods for analysis of ignorable multilevel missing data.

2 Missing Data Methods

A wide array of methods exist for efficient analysis of single-level ignorable missing data (Rubin 1976, 1987, 1996; Dempster et al. 1977; Meng 1994; Schafer 1997, 2003; Little and Rubin 2002). In particular, model-based multiple imputation (Rubin 1987, 1996) is now routinely available based on widely used software packages such as NORM (Schafer 1999) and SAS (PROC MI, Y.C. Yuan 2000). These single-level methods, however, cannot be applied validly to hierarchical missing data and their extension to multilevel data entails challenges (Dempster et al. 1981; Schafer and Yucel 2002; Goldstein and Browne 2002, 2005; Yucel 2008; Shin and Raudenbush 2007, 2010, 2011). If methods developed for the multiple imputation of single-level data are applied to multilevel data, the variance-covariance structure of the imputed data sets will not accurately represent the multilevel educational processes that generated the data, nor will the structural relations at each level be captured correctly. When multilevel data are analyzed by a single-level method or under the misspecified number of levels, the resulting inferences may be considerably biased or inefficient (Shin 2003; Shin and Raudenbush 2011; Van Buuren 2011).

Current widely available methods for efficient analysis of ignorable multilevel missing data are quite limited. A two-level multivariate hierarchical linear model, where level-1 outcomes are subject to missingness given completely observed covariates, may be efficiently estimated via software packages such as Mplus (Muthen and Muthen 2010) and MLwiN (Rasbash et. al. 2009; Browne 2012). With a univariate outcome in the model, this approach amounts to the complete-case analysis. When outcomes and covariates have missing values in the hierarchical model, however, a joint distribution of the variables subject to missingness has to be formulated and estimated to efficiently handle the missing data; and given the estimated distribution, multiple imputation of the complete data may be generated for subsequent analysis of the desired hierarchical model (Rubin 1987). Software packages such as WinBUGS (Spiegelhalter et al. 2000; Lunn et al. 2009), Mplus (Muthen and Muthen 2010; Asparouhov and Methen 2010), MLwiN (Browne 2012) and R (Yucel 2008) provide Bayesian methods that enable formulation and estimation of such a joint distribution, and generation of the multiple imputation for subsequent analysis of the hierarchical model. However, these software packages provide little guidance as to how to explicitly formulate the joint distribution corresponding to the hierarchical model. For example, formulation of the joint distribution given a level-2 covariate subject to missingness in the hierarchical model is neither automated nor clearly described by any of the software packages. In general, the transformation between the joint distribution and the hierarchical model is nontrivial, involving an identification problem, and great care should be taken in formulation of the joint distribution that will identify the hierarchical model (Meng 1994; Shin and Raudenbush 2007). Otherwise, the estimation may produce biased point and uncertainty estimates of the hierarchical model or the formulated joint distribution may be extremely high-dimensional to estimate well (Shin and Raudenbush 2007, 2013).

Multilevel ignorable missing data may be multiply imputed by univariate sequential regression models (Raghunathan et al. 2001), which is also known as multiple imputation by fully conditional specification (Van Buuren et al. 2006). Software packages such as IVEware (Raghunathan et al. 2001) and Multivariate Imputation by Chained Equations (MICE, van Buuren and Groothuis-Oudshoorn 2011) use a Bayesian method to produce multiple imputation. This approach specifies a univariate regression model for each variable subject to missingness conditional on all other variables and generates multiple imputation based on the fitted model. While flexible in dealing with a mixture of continuous and discrete variables subject to missingness, the chained univariate models may not be compatible with a joint model (Horton and Kleinman 2007; van Buuren and Groothuis-Oudshoorn 2011). The implied joint model by the series of univariate regression models may not exist (Rubin 2003; Van Buuren et al. 2006). This approach has not been extended to outcomes and covariates subject to missingness at multiple levels of a hierarchical model (Van Buuren 2011).

The next section introduces an efficient missing data method via multiple imputation and its software for unbiased and efficient estimation of a two-level hierarchical linear model given ignorable missing data. A key feature is that the data analyst need only know the desired hierarchical model. This approach removes or substantially reduces the burden of the incomplete data analysis from the data analyst as intended by the method of multiple imputation (Rubin 1987, 1996; Meng 1994). Consequently, with the software in hand, the incomplete hierarchical data analysis will introduce little more challenge than the complete-data counterpart to the data analyst.

3 Efficient Handling of Missing Data

This section explains how to efficiently estimate a two-level hierarchical linear model (HLM2) given incomplete data according to the missing data method of Shin and Raudenbush (2007). The method employs a six-step analysis procedure to: (1) specify a desired hierarchical linear model given incompletely observed hierarchical data; (2) reparametrize as the joint distribution of variables, including the outcome, that are subject to missingness conditional on all of the covariates that are completely observed under multivariate normality; (3) efficiently estimate the joint distribution using maximum likelihood (ML); (4) generate multiple imputation of complete data based on the ML estimates of the joint model; (5) analyze the desired hierarchical model by complete-data analysis given the multiple imputation; and finally (6) combine the multiple hierarchical model estimates (Rubin 1987). These steps have been implemented in a software package *HLM 7* (Raudenbush et al. 2011). Given the hierarchical linear model that a data analyst specifies at the first step, *HLM 7* automates the rest of the analysis procedure to produce efficient analysis of the hierarchical model. Consequently, the data analyst need only know her desired hierarchical linear model which is no different from the complete-data analysis.

In this section, we introduce two comparatively simple examples of hierarchical linear models with incomplete data to describe the problem that researchers confront in the conventional incomplete data analysis and how the missing data method resolves the problem by enabling efficient analysis via multiple imputation. One is a random-intercept model, and the other a random-coefficients model. Then, we present a reasonably general HLM2 given incomplete data, which may be efficiently analyzed by the method. Finally, we describe how to estimate the desired parameters and make inferences given multiple imputation (Rubin 1987).

3.1 Random-Intercept Model

To see how to handle multilevel incomplete data efficiently, it is instructive to consider a simple random-intercept model (Raudenbush and Bryk 2002). Let child i attend school j for $i = 1, \cdots, n_j$ and $j = 1, \cdots, M$. We consider a simple child-level or level-1 model

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

where Y_{ij} is a univariate outcome variable, β_{0j} is the level-1 intercept, β_{1j} is the effect of a level-1 covariate X_{ij} and child-specific random effect ϵ_{ij} is normally distributed with mean zero and variance σ^2, i.e., $\epsilon_{ij} \sim N(0, \sigma^2)$. Note that the model (1) is single-level within school j. Each coefficient in the level-1 model (1) becomes an outcome variable that may vary across schools in the school-level or level-2 model. We consider level-2 models

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01} Z_j + u_{0j}, \\ \beta_{1j} &= \gamma_{10} \end{aligned} \tag{2}$$

where γ_{00}, γ_{01} and γ_{10} are level-2 coefficients, Z_j is a level-2 covariate, and school-specific random effect $u_{0j} \sim N(0, \tau)$ is independent of child-specific ϵ_{ij}. By replacing β_{0j} and β_{1j} in the level-1 model (1) with $\gamma_{00} + \gamma_{01} Z_j + u_{0j}$ and γ_{10} on the right-hand side of the level-2 models, respectively, we obtain a random-intercept model or HLM2

$$Y_{ij} = \gamma_{00} + \gamma_{01} Z_j + \gamma_{10} X_{ij} + u_{0j} + \epsilon_{ij}. \tag{3}$$

With data completely observed, this model may be analyzed by standard multilevel software such as SAS, *HLM 7* and MLwiN (Rasbash et. al. 2009).

Difficulties arise given incompletely observed data. We consider (Y_{ij}, X_{ij}, Z_j) all subject to missingness with a general missing pattern in the desired model (3). Missing data may occur under seven different patterns for child i attending school j: One, two or all three values of (Y_{ij}, X_{ij}, Z_j) may be missing. In general, p variables subject to missingness may produce up to $2^p - 1$ different missing patterns. Complete-case analysis drops children or observations that belong to any one of the missing patterns. It also deletes school j with missing Z_j which entails deletion of all students attending the school. The resulting inferences will be inefficient and subject to bias. Ad hoc single-imputation methods fill in missing values for subsequent complete-data analysis. The imputed data under-represent uncertainty due to missing data in estimation. In general, hypothesis tests will be liberal, rejecting the null hypothesis too often. These methods are not recommended unless children and schools with missing values consist of a small fraction of all children and schools.

Efficient analysis of the model (3) has to analyze all available data. That is, rather than dropping observations that belong to any one of the seven missing patterns in the complete-case analysis, we drop child i in school j if and only if she belongs to one missing pattern: all three values of (Y_{ij}, X_{ij}, Z_j) missing. If one of (Y_{ij}, X_{ij}, Z_j) is missing for the child, the other two values available are analyzed; and if two values out of (Y_{ij}, X_{ij}, Z_j) are missing, the one value observed is analyzed. Consequently, children with unit non-response are also analyzed as long as they attend schools having observed Z_j to strengthen inferences at school level. Consider, for example, school j having a single child sampled ($n_j = 1$) who misses both Y_{ij} and X_{ij}, but school j has Z_j observed. Note that school j is dropped if and only if all school mates miss both Y_{ij} and X_{ij} and the school misses Z_j.

Efficient analysis of the HLM2 (3) using all available data may be formalized in the joint distribution of (Y_{ij}, X_{ij}, Z_j) subject to missingness

$$\begin{bmatrix} Y_{ij} \\ X_{ij} \\ Z_j \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} b_{1j} \\ b_{2j} \\ b_{3j} \end{bmatrix} + \begin{bmatrix} e_{1ij} \\ e_{2ij} \\ 0 \end{bmatrix} \tag{4}$$

for the means $(\alpha_1, \alpha_2, \alpha_3)$ of (Y_{ij}, X_{ij}, Z_j), and school-specific random effects $\begin{bmatrix} b_{1j} \\ b_{2j} \\ b_{3j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \psi_{11} & \psi_{12} & \psi_{13} \\ \psi_{12} & \psi_{22} & \psi_{23} \\ \psi_{13} & \psi_{23} & \psi_{33} \end{bmatrix}\right)$ of (Y_{ij}, X_{ij}, Z_j) independent of child-specific random effects $\begin{bmatrix} e_{1ij} \\ e_{2ij} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}\right)$ of (Y_{ij}, X_{ij}) where $\psi_{11} = var(b_{1j}), \psi_{12} = cov(b_{1j}, b_{2j}), \psi_{13} = cov(b_{1j}, b_{3j}), \psi_{22} = var(b_{2j}), \psi_{23} = cov(b_{2j}, b_{3j}), \psi_{33} = var(b_{3j}), \sigma_{11} = var(e_{1ij}), \sigma_{12} = cov(e_{1ij}, e_{2ij})$ and $\sigma_{22} =$

$var(e_{2ij})$. Note that Z_j stays the same among schoolmates with no level-1 random effect. The missing data method for the desired hierarchical model (3) via efficient estimation of the joint model (4) produces efficient analysis of the hierarchical model as the conditional distribution of Y_{ij} given X_{ij} and Z_j (Shin and Raudenbush 2007).

To explicitly show how to analyze all available data, we first consider children with a single value missing. If a single value Y_{ij} is missing for child i attending school j, the two observed values (X_{ij}, Z_j) of the child enable estimation of

$$\begin{bmatrix} X_{ij} \\ Z_j \end{bmatrix} = \begin{bmatrix} \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} b_{2j} \\ b_{3j} \end{bmatrix} + \begin{bmatrix} e_{2ij} \\ 0 \end{bmatrix} \sim N\left(\begin{bmatrix} \alpha_2 \\ \alpha_3 \end{bmatrix}, \begin{bmatrix} \psi_{22}+\sigma_{22} & \psi_{23} \\ \psi_{23} & \psi_{33} \end{bmatrix} \right) \quad (5)$$

which adds information to estimation for $(\alpha_2, \alpha_3, \psi_{22}, \psi_{23}, \psi_{33}, \sigma_{22})$; if a single value X_{ij} is missing, the other two observed values enable estimation of a bivariate model (Y_{ij}, Z_j) to strengthen inferences involving $(\alpha_1, \alpha_3, \psi_{11}, \psi_{13}, \psi_{33}, \sigma_{11})$; and with Z_j missing, child i with observed (Y_{ij}, X_{ij}) adds information to estimation for $(\alpha_1, \alpha_2, \psi_{11}, \psi_{12}, \psi_{22}, \sigma_{11}, \sigma_{12}, \sigma_{22})$ in a bivariate model (Y_{ij}, X_{ij}).

Let us now consider children with two values missing. Child i missing (Y_{ij}, X_{ij}) adds information to estimation of a univariate model $Z_j \sim N(\alpha_3, \psi_{33})$ at school level. Take, for example, school j having a single child ($n_j = 1$) with unit nonresponse, but the school has observed Z_j. If the child misses (Y_{ij}, Z_j), she contributes to estimation of $X_{ij} \sim N(\alpha_2, \psi_{22} + \sigma_{22})$; and if she misses (X_{ij}, Z_j), she adds information to estimation of $Y_{ij} \sim N(\alpha_1, \psi_{11} + \sigma_{11})$.

Consequently, all partially observed cases contribute to estimation of the joint model (4), and thus of the desired model (3). The only case when school j is dropped from analysis happens if and only if school j misses Z_j and all n_j schoolmates miss both Y_{ij} and X_{ij}. Therefore, the method analyzes all available data to achieve efficient analysis of the desired hierarchical model (3).

Completely Observed Covariates. We now consider completely observed covariates U_{ij} and W_j at levels 1 and 2, respectively, in addition to (Y_{ij}, X_{ij}, Z_j) subject to missingness. The desired level-1 model is

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + \beta_{2j}U_{ij} + \epsilon_{ij} \tag{6}$$

where β_{2j} is the effect of the level-1 covariate U_{ij} and everything else is defined in the same way as that of the model (1). We consider level-2 models

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01}Z_j + \gamma_{02}W_j + u_{0j}, \qquad (7)\\ \beta_{1j} &= \gamma_{10}, \\ \beta_{2j} &= \gamma_{20} \end{aligned}$$

where γ_{00}, γ_{01}, γ_{02}, γ_{10} and γ_{20} are level-2 coefficients, Z_j and W_j are level-2 covariates, and school-specific random effect $u_{0j} \sim N(0, \tau)$ is independent of child-specific $\epsilon_{ij} \sim N(0, \sigma^2)$. The desired random-intercept model or HLM2 is

$$Y_{ij} = \gamma_{00} + \gamma_{01}Z_j + \gamma_{02}W_j + \gamma_{10}X_{ij} + \gamma_{20}U_{ij} + u_{0j} + \epsilon_{ij}. \tag{8}$$

To efficiently handle missing data, we formulate the joint distribution of (Y_{ij}, X_{ij}, Z_j) subject to missingness conditional on (U_{ij}, W_j) completely observed. That is, we formulate the joint model as Equation (4) where α_1, α_2 and α_3 are replaced with $\alpha_{10} + \alpha_{11}W_j + \alpha_{12}U_{ij}$, $\alpha_{20} + \alpha_{21}W_j + \alpha_{22}U_{ij}$ and $\alpha_{30} + \alpha_{31}W_j$, respectively, and every other component is the same as it appears in the model (4). Note that the level-2 covariate W_j has its effects on level-1 as well as level-2 responses (Y_{ij}, X_{ij}, Z_j) while the level-1 covariate U_{ij} affects level-1 responses (Y_{ij}, X_{ij}) only. The efficient handling of missing data for the joint model (4) also applies here for the joint model corresponding to the HLM2 (8).

3.2 Random-Coefficients Model

This section explains the strategy for efficient analysis of a random-coefficients model given incomplete data. We consider the level-1 model (6) where the intercept as well as the coefficient of U_{ij} vary randomly across schools at level 2. Thus, we consider level-2 models

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01} Z_j + \gamma_{02} W_j + u_{0j}, \\ \beta_{1j} &= \gamma_{10}, \\ \beta_{2j} &= \gamma_{20} + u_{2j} \end{aligned} \tag{9}$$

where school-specific random effects $u_j \sim N(0, \tau)$ are independent of child-specific $\epsilon_{ij} \sim N(0, \sigma^2)$ for $u_j = \begin{bmatrix} u_{0j} \\ u_{2j} \end{bmatrix}$ and $\tau = \begin{bmatrix} \tau_{00} & \tau_{02} \\ \tau_{02} & \tau_{22} \end{bmatrix}$ and everything else is defined identically as the counterpart of the level-2 models (7). The desired random-coefficients model or HLM2 is

$$Y_{ij} = \gamma_{00} + \gamma_{01} Z_j + \gamma_{02} W_j + \gamma_{10} X_{ij} + \gamma_{20} U_{ij} + u_{0j} + u_{2j} U_{ij} + \epsilon_{ij} \tag{10}$$

for (Y_{ij}, X_{ij}, Z_j) subject to missingness, and (U_{ij}, W_j) completely observed. Conventional analysis of the HLM2 (10) confronts the same problems with the seven missing data patterns as does that of the random-intercept model (3). Efficient handling of missing data for the hierarchical model (3) also applies here for efficient analysis of the hierarchical model (10). We consider the joint distribution of (Y_{ij}, X_{ij}, Z_j) subject to missingness conditional on completely observed U_{ij} and W_j as

$$\begin{bmatrix} Y_{ij} \\ X_{ij} \\ Z_j \end{bmatrix} = \begin{bmatrix} \alpha_{10} + \alpha_{11} W_j + \alpha_{12} U_{ij} \\ \alpha_{20} + \alpha_{21} W_j + \alpha_{22} U_{ij} \\ \alpha_{30} + \alpha_{31} W_j \end{bmatrix} + \begin{bmatrix} b_{0j} + b_{1j} U_{ij} \\ b_{2j} \\ b_{3j} \end{bmatrix} + \begin{bmatrix} e_{1ij} \\ e_{2ij} \\ 0 \end{bmatrix} \tag{11}$$

where $(\alpha_{10}, \alpha_{20}, \alpha_{30})$ are the intercepts, $(\alpha_{11}, \alpha_{21}, \alpha_{31})$ are the effects of W_j on (Y_{ij}, X_{ij}, Z_j), respectively, $(\alpha_{12}, \alpha_{22})$ are the effects of U_{ij} on (Y_{ij}, X_{ij}), respectively, and school-specific $\begin{bmatrix} b_{0j} \\ b_{1j} \\ b_{2j} \\ b_{3j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \psi_{00} & \psi_{01} & \psi_{02} & \psi_{03} \\ \psi_{01} & \psi_{11} & \psi_{12} & \psi_{13} \\ \psi_{02} & \psi_{12} & \psi_{22} & \psi_{23} \\ \psi_{03} & \psi_{13} & \psi_{23} & \psi_{33} \end{bmatrix} \right)$ are independent of child-specific $\begin{bmatrix} e_{1ij} \\ e_{2ij} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} \right)$.

Covariates Subject to Missingness Having Random Effects. Efficient estimation of the random-coefficients model (10) requires that covariate U_{ij} having random coefficient u_{2j} be completely observed. Difficulty arises when U_{ij} is subject to missingness in the hierarchical model. The joint model for $(Y_{ij}, X_{ij}, U_{ij}, Z_j)$ subject to missingness has to be formulated for efficient handling of missing data while U_{ij} needs to be given on the right hand side of the model for estimation of its random coefficient. Such a joint model cannot be expressed as a multivariate normal distribution so that the normal factorization of the joint model that leads to the hierarchical model as the conditional distribution of Y_{ij} given covariates does not apply. Consequently, it is difficult to efficiently handle missing data in the hierarchical model via ML estimation of the multivariate normal joint model. We assume that covariates having random effects are completely observed, which is a limitation of the method.

3.3 General HLM2

We now express a general HLM2 given incomplete data, which can be efficiently analyzed by the missing data method. The model is

$$Y_{ij} = X_{ij}^T\gamma_x + Z_j^T\gamma_z + U_{ij}^T\gamma_u + W_j^T\gamma_w + D_{ij}^T u_j + \epsilon_{ij} \tag{12}$$

where Y_{ij} is a univariate outcome variable, X_{ij} and Z_j are vectors of p_1 level-1 and p_2 level-2 covariates subject to missingness having fixed effects γ_x and γ_z, respectively, U_{ij} and W_j are vectors of p_3 level-1 and p_4 level-2 covariates completely observed having fixed effects γ_u and γ_w, respectively, and D_{ij} is another vector of p_5 level-1 covariates completely observed having level-2 unit-specific random effects $u_j \sim N(0,\tau)$ independent of level-1 unit-specific random errors $\epsilon_{ij} \sim N(0,\sigma^2)$ for a p_5-by-p_5 matrix τ and scalar σ^2. The desired parameters are $\theta = (\gamma_x, \gamma_z, \gamma_u, \gamma_w, \tau, \sigma^2)$.

The hierarchical models considered so far are special cases of the HLM2 (12). For example, the random-intercept model (3) is a special case of the HLM2 (12) where X_{ij} and Z_j are scalar, $U_{ij} = 0$, $W_j = D_{ij} = 1$, $\gamma_x = \gamma_{10}$, $\gamma_z = \gamma_{01}$, $\gamma_u = 0$, $\gamma_w = \gamma_{00}$, and $u_j = u_{0j}$. Although the general HLM2 (12) is not required to represent an intercept, many applications do where the first elements of W_j and D_{ij} are equal to one with the corresponding first elements of γ_w and u_j representing the mean intercept and the random deviation of the intercept from the mean, respectively. We require that D_{ij} be completely observed. Note that U_{ij} and D_{ij} may share common covariates. For example, D_{ij} may be a subset of covariates in U_{ij}.

The p variables (Y_{ij}, X_{ij}, Z_j) subject to missingness in the HLM2 (12) may produce up to $2^p - 1$ different missing patterns for $p = 1 + p_1 + p_2$. Complete-case analysis drops children who belong to any one of the missing patterns. As the number of p variables subject to missingness increases, a number of children and schools may have to be dropped from the analysis which results in inefficient inferences that may also be substantially biased.

To describe a joint model that efficiently handles missing data in the HLM2 (12), let $A \otimes B$ be a Kronecker product that multiplies a b-by-b matrix B to each element of an a-by-a matrix A (Magus and Neudecker 1988), and let I_n denote an n-by-n identity matrix for a positive integer n. For example, $I_3 \otimes B$ is a $(3 \times b)$-by-$(3 \times b)$ diagonal matrix with diagonal submatrices (B, B, B) and all other elements equal to zero. Given the HLM2 (12) with missing data, we formulate the joint distribution of (Y_{ij}, X_{ij}, Z_j) subject to missingness conditional on (U_{ij}, W_j, D_{ij}) completely observed as

$$\begin{bmatrix} Y_{ij} \\ X_{ij} \\ Z_j \end{bmatrix} = \begin{bmatrix} U_{ij}^{*T}\alpha_1 \\ (I_{p_1} \otimes U_{ij}^{*T})\alpha_2 \\ (I_{p_2} \otimes W_j^T)\alpha_3 \end{bmatrix} \begin{bmatrix} D_{ij}^T b_{1j} \\ b_{2j} \\ b_{3j} \end{bmatrix} + \begin{bmatrix} e_{1ij} \\ e_{2ij} \\ 0 \end{bmatrix} \tag{13}$$

for $U_{ij}^* = [W_j^T \; U_{ij}^T]^T$, vectors α_1 and α_2 of the fixed effects of U_{ij}^* on Y_{ij} and X_{ij}, respectively, a vector α_3 of the fixed effects of W_j on Z_j, and school-specific random effects $\begin{bmatrix} b_{1j} \\ b_{2j} \\ b_{3j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \psi_{11} & \psi_{12} & \psi_{13} \\ \psi_{12} & \psi_{22} & \psi_{23} \\ \psi_{13} & \psi_{23} & \psi_{33} \end{bmatrix}\right)$ and child-specific random effects $\begin{bmatrix} e_{1ij} \\ e_{2ij} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}\right)$ independent.

The missing data method for the HLM2 (12) via efficient estimation of the joint model produces efficient analysis of the hierarchical model as the conditional distribution of Y_{ij} given covariates (Shin and Raudenbush 2007). Note that given complete data, the HLM2 (12) is $Y_{ij} = U_{ij}^T\gamma_u + W_j^T\gamma_w + D_{ij}^T u_j + \epsilon_{ij}$ equal to the joint model (13). The same strategy described above is used for efficient handling of missing data in (Y_{ij}, X_{ij}, Z_j) where child i with at least a single value observed contributes to estimation of the joint model. See Shin and Raudenbush (2007) for ML estimation of the joint model and multiple imputation given the ML estimates. Note that the variables subject to missingness, including the outcome, appear on the left-hand side given those completely observed on the right-hand side, which is the required form of the joint model (13) for efficient handling of missing data and efficient computation.

3.4 Combining Estimates from Multiple Imputation

Analysis of each of m imputed or completed data sets according to the desired HLM2 (12) produces m sets of θ estimates and their associated variances. Following Rubin (1987) and Schafer (1997), let Q be a parameter or a function of parameters in θ. Analysis of the tth completed data set produces the ML estimate $\hat{Q}_t$ and the associated variance U_t for $t = 1, 2, \cdots, m$. The combined parameter estimate is simply the average

$$\overline{Q} = \frac{1}{m}\sum_{t=1}^{m}\hat{Q}_t. \tag{14}$$

The variance associated with the combined estimate is

$$T = \overline{U} + \left(1 + \frac{1}{m}\right)B \tag{15}$$

that consists of the average within-imputation variance

$$\overline{U} = \frac{1}{m}\sum_{t=1}^{m}U_t \tag{16}$$

and the between-imputation variance

$$B = \frac{1}{m-1}\sum_{t=1}^{m}(\hat{Q}_t - \overline{Q})^2. \tag{17}$$

The within-imputation variance U_t reflects uncertainty in estimation of Q given the tth imputed data set (as if the missing values imputed were the true values) while the between-imputation variance B conveys uncertainty across the m estimates of Q due to missing data. No missing data implies $B = 0$ so that $T = \overline{U}$. With the infinite number of imputations, the variance (15) associated with $\overline{Q}$ is $T = \overline{U} + B$. The term $(1 + 1/m)$ in Equation (15) adds extra uncertainty due to the finite number of m imputations. (Rubin 1987; Schafer 1997; Little and Rubin 2002)

For inferences on a column vector Q of k elements, Equations (14) to (16) are of the same form, but Equation (17) becomes $B = \frac{1}{m-1}\sum_{t=1}^{m}(\hat{Q}_t - \overline{Q})(\hat{Q}_t - \overline{Q})^T$ where $(\hat{Q}_t - \overline{Q})^T$ denotes the vector $(\hat{Q}_t - \overline{Q})$ transposed.

3.5 Hypothesis Tests

Let Q be a fixed effect or a linear function of fixed effects. We make inferences about Q based on

$$\frac{\overline{Q} - Q}{\sqrt{T}} \quad \sim \quad t_\nu \tag{18}$$

where t_ν is the t distribution with the degrees of freedom

$$\nu \quad = \quad (m-1)\left(1 + \frac{1}{r}\right)^2 \tag{19}$$

for

$$r \quad = \quad \left(1 + \frac{1}{m}\right)\frac{B}{\overline{\overline{U}}} \tag{20}$$

estimating "the relative increase in variance due to nonresponse" (Rubin 1987). Consequently, a $(1-\alpha) \times 100\%$ confidence interval for Q is

$$\overline{Q} \pm t_{\nu,1-\alpha/2}\sqrt{T} \tag{21}$$

where $t_{\nu,1-\alpha/2}$ is the $(1-\alpha/2)\times 100$th percentile from t_ν. The p-value for testing a null hypothesis $H_0 : Q = Q_0$ against an alternative hypothesis $H_a : Q \neq Q_0$ at a significance level α is

$$2 \times P\left(\mathcal{T} > \frac{|\overline{Q} - Q_0|}{\sqrt{T}}\right) \tag{22}$$

where $\mathcal{T}$ is a t_ν random variable (Rubin 1987; Schafer 1997).

When the between-imputation variance B is low relative to $\overline{U}$ to yield a low r such that the degrees of freedom in Equation (19) are high, $t_{\nu,1-\alpha/2}$ in the interval (21) and $\mathcal{T}$ in the p-value (22) can be replaced with the corresponding percentile $z_{1-\alpha/2}$ and a random variable Z from the standard normal distribution, respectively. When the relative increase r in variance due to missing data is high to yield low degrees of freedom, Equation (20) implies that increasing the number of m imputations decreases r to raise the degrees of freedom. *HLM 7* prints the degrees of freedom ν in Equation (19) that can be solved for

$$r \quad = \quad \left(\sqrt{\nu/(m-1)} - 1\right)^{-1}. \tag{23}$$

Equation (20) with $\overline{U}$ replaced with T estimates "the fraction of information about Q missing due to nonresponse" (Little and Rubin 2002)

$$s \quad = \quad \left(1 + \frac{1}{m}\right)\frac{B}{T}. \tag{24}$$

Both r and s are positively associated with the between-imputation variance

B, but negatively associated with the number of m imputations. Frequently, researchers claim that a few imputations are enough to handle missing data reasonably well based on the fraction of missing information estimated by Equation (24). The size of the standard error of the Q estimate based on m imputations relative to the ideal one based on infinitely many imputations is approximately $\sqrt{1+s/m}$ (Rubin 1987, p. 114; Schafer 1997, p. 107). When the fraction of missing information is high at $s = 0.5$, for example, the standard error $\sqrt{T}$ of $\overline{Q}$ based on $m = 5$ imputations will be about $\sqrt{1+0.5/5} = 1.05$ times as high as the counterpart based on infinitely many imputations. With $s = 0.3$, as few as $m = 3$ imputations will achieve about the same efficiency in terms of the relative size of standard errors.

To compare model fits based on the likelihood ratio tests, let Q be k parameters in θ of HLM2 (12) or the full model. We want to test a null hypothesis $H_0 : Q = Q_0$ versus an alternative one $H_a : Q \neq Q_0$. Let θ_0 be the parameters of the reduced model under $Q = Q_0$. Consider, for example, a 2-dimensional $Q = \begin{bmatrix} \tau_{01} \\ \tau_{11} \end{bmatrix}$ in $\theta = (\gamma_x, \gamma_z, \gamma_u, \gamma_w, \tau, \sigma^2)$ for $\tau = \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{01} & \tau_{11} \end{bmatrix}$ and $Q_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. Then, θ_0 has $k = 2$ parameters less than θ. The Q may also include a combination of fixed effects, variances and covariances.

Let $\hat{\theta}^t$ and $\hat{\theta}_0^t$ be the ML estimates of θ and θ_0, respectively, given the tth completed data set for $t = 1, \cdots, m$. The log likelihoods $l(\hat{\theta}^t)$ and $l(\hat{\theta}_0^t)$ evaluated at $\hat{\theta}^t$ and $\hat{\theta}_0^t$, respectively, yield the likelihood ratio statistic $d_t = 2[l(\hat{\theta}^t) - l(\hat{\theta}_0^t)]$. The test statistic proposed by Li et al. (1991) is

$$D_1 = \frac{\bar{d}/k - (m+1)(m-1)^{-1} r_1}{1 + r_1} \tag{25}$$

for $\bar{d} = \sum_{t=1}^{m} d_t/m$ where

$$r_1 = \left(1 + \frac{1}{m}\right)\left[\frac{1}{m-1}\sum_{t=1}^{m}\left(\sqrt{d_t} - \overline{\sqrt{d}}\right)^2\right] \qquad (26)$$

is $(1+1/m)$ times the sample variance of $\sqrt{d_1}, \cdots, \sqrt{d_m}$ for $\overline{\sqrt{d}} = \sum_{t=1}^{m} \sqrt{d_t}/m$ (Little and Rubin 2002). The r_1 estimates the average relative increase in variance due to missing data across the k parameters Q (Schafer 1997). Let F_{k,ν_1} denote a random variable from the F distribution with k numerator and ν_1 denominator degrees of freedom for

$$\nu_1 = k^{-3/m}(m-1)(1+1/r_1)^2. \qquad (27)$$

The p-value is given by

$$P(F_{k,\nu_1} > D_1). \qquad (28)$$

With r_1 close to zero, ν_1 is large so that kD_1 has the chi-square distribution with k degrees of freedom for $D_1 \approx \bar{d}/k$. Then the p-value (28) may also be obtained by

$$P(\chi_k^2 > kD_1) \qquad (29)$$

for a chi-square random variable χ_k^2 with k degrees of freedom. Given multiple imputation, the likelihood ratio statistic d_t may be obtained from the tth completed data set to yield r_1, D_1 and ν_1 for the hypothesis test. The test statistic D_1 yields an approximate range of p-values between one half and twice the computed p-value (Li et al. 1991).

To obtain a more accurate p-value, let $\bar{\theta} = \sum_{t=1}^{m} \hat{\theta}^t/m$ and $\bar{\theta}_0 = \sum_{t=1}^{m} \hat{\theta}_0^t/m$ so that $d_t' = 2[l_t(\bar{\theta}) - l_t(\bar{\theta}_0)]$ is the likelihood ratio test statistic evaluated at the average ML estimates $\bar{\theta}$ and $\bar{\theta}_0$ given the tth completed data set. The test statistic proposed by Meng and Rubin (1992) is

$$D_2 = \frac{\bar{d}'}{k(1+r_2)} \tag{30}$$

for $\bar{d}' = \sum_{t=1}^{m} d_t'/m$ where

$$r_2 = \frac{(m+1)}{k(m-1)}(\bar{d} - \bar{d}') \tag{31}$$

estimates the average relative increase in variance due to missing data across the k parameters Q (Schafer 1997). The p-value is given by

$$P(F_{k,\nu_2} > D_2) \tag{32}$$

where the denominator degrees of freedom is

$$\nu_2 = \begin{cases} 4 + (u-4)[1 + (1-2/u)/r_2]^2, & \text{if } u = k(m-1) > 4 \\ (m-1)(k+1)(1+1/r_2)^2/2, & \text{otherwise.} \end{cases} \tag{33}$$

Unlike D_1, computation of D_2 requires log likelihoods $l_t(\bar{\theta})$ and $l_t(\bar{\theta}_0)$ evaluated at the average ML estimates $\bar{\theta}$ and $\bar{\theta}_0$ given the tth completed data set that $HLM7$ does not provide at the time of my writing this chapter.

The two approaches to testing $H_0 : Q = Q_0$ against $H_a : Q \neq Q_0$ (Li et al. 1991; Meng and Rubin 1992) are based on the likelihood ratio statistics. When Q_0 involves variance components equal to zero given complete data, the likelihood ratio test is known to produce a conservative p-value based on the chi-square distribution with k degrees of freedom (Pinheiro and Bates 2000). Stram and Lee (1994) suggested use of a mixture of chi-square distributions to improve the accuracy of the p-value (Pinheiro and Bates 2000; Verbeke and

Molenberghs 2000; Snijders and Bosker 2012). Given incomplete data, the test statistic D_1 yields an approximate range of p-values between one half and twice the observed p-value (Li et al. 1991), and the test statistic D_2 produces the p-value (32) that is more accurate than the corresponding p-value (28) (Meng and Rubin 1992; Schafer 1997; Little and Rubin 2002). The next two sections show how to analyze a hierarchical linear model given ignorable missing data by *HLM 7* according to the method explained in this section.

4 Data Analysis

This section illustrates how to efficiently analyze hierarchical linear model (12) given incompletely observed data from the Programme for International Student Assessment (PISA, OECD 2007). PISA has been collecting hierarchical data about 15-year old students attending schools nested within nations every three years since the year 2000. The data for analysis consists of 5611 students attending 166 schools in the US from the PISA 2006 data collection. Table 1 summarizes the data. The outcome variable of interest is the mathematics achievement score (MATH). PISA imputes each missing score five times to provide five sets of plausible mathematics scores. In this section, we analyze the first set of plausible mathematics scores summarized in Table 1 as if they were completely observed. The next section illustrates an analysis strategy with all plausible values.

To summarize the data for analysis, at level 1, mathematics score (MATH) and age (AGE) are completely observed while the highest parental occupation status (HISEI), the highest education level of parents in the number of years of schooling (PARENTED), family wealth (WEALTH) and first-generation immigrant status (IMMIG1) are missing for 390, 61, 34, and 189 students, respectively. The 5611 students score 475 points in mathematics and are 190 months old on average; the highest occupation status and education level of parents are 52.46 units and 13.61 years on average, respectively; the average family wealth is 0.15 units; and 6% of the students are first-generation immigrants. At level 2, the student-to-teacher ratio (STRATIO) is missing for 28 schools, or 17% of the 166 schools, and the private school indicator (PRIVATE) is missing for 3 schools. The schools have 15.46 students per teacher on average, and 9% of the schools are private (cf. OECD 2007).

Summary statistics reveal that first-generation immigrants scored 36.08 points lower than did other students in mathematics achievement on average. In this section, we ask how much of the difference is attributable to the individual and school characteristics summarized in Table 1; and, controlling for the individual and school characteristics, how first-generation immigrants compare with other students in mathematics achievement. The complete-case analysis drops 1405 students and 28 schools to produce inefficient inferences that may also be substantially biased. We compare the complete-case analysis with the efficient missing data analysis given incomplete data.

4.1 Complete-Case Analysis

Preliminary analysis reveals that the school means of the highest parent education (PARENTED), indicative of school quality, vary substantially across schools with a 95% confidence interval (11.38, 15.82). The effect of the highest parent education may vary randomly across schools of different quality. A random-coefficients model to test such a hypothesis is

$$\begin{aligned} Y_{ij} &= \gamma_{00} + \gamma_{01} STRATIO + \gamma_{02} PRIVATE + \gamma_{10} AGE + \gamma_{20} HISEI \\ &\quad + \gamma_{30} PARENTED + \gamma_{40} WEALTH + \gamma_{50} IMMIG1 \\ &\quad + u_{0j} + u_{3j} PARENTED + \epsilon_{ij}, \end{aligned} \tag{34}$$

a special case of the HLM2 (12) where

$$Y_{ij} = MATH,\, X_{ij} = [HISEI\ PARENTED\ WEALTH\ IMMIG1],$$
$$Z_j = [STRATIO\ PRIVATE],\, U_{ij} = AGE,\, W_j = 1,\, D_{ij} = [1\ PARENTED],$$
$$\gamma_x = [\gamma_{20}\ \gamma_{30}\ \gamma_{40}\ \gamma_{50}]^T,\, \gamma_z = [\gamma_{01}\ \gamma_{02}]^T,\, \gamma_u = \gamma_{10},\, \gamma_w = \gamma_{00},\, u_j = [u_{0j}\ u_{3j}]^T$$

for $\tau = \begin{bmatrix} \tau_{00} & \tau_{03} \\ \tau_{03} & \tau_{33} \end{bmatrix}$. We center HISEI, PARENTED, WEALTH, AGE and STRATIO around their respective sample means, and carry out the complete-case analysis by *HLM 7* to produce the ML estimates under the heading "CC" in Table 2. The CC analysis analyzed 4206 students attending 138 schools. The τ_{33} estimate is 19 with the associated variance estimate 6.45^2, not shown in Table 2. The hypothesis test of interest is

$$H_0 : \tau_{33} = 0 \text{ against } H_a : \tau_{33} > 0.$$

Approximate normality of the ML estimator $ln(\hat{\tau}_{33})$ for the natural logorithm $ln(\cdot)$ produces an approximate 95% confidence interval for $ln(\tau_{33})$ which is transformed to an approximate 95% confidence interval $(9.70, 36.90)$ for τ_{33}. The interval far away from zero provides some evidence in support of the alternative hypothesis. For the hypothesis test, *HLM 7* produces a χ^2 test statistic 196.47 with 136 degrees of freedom based on 137 schools with enough data (Raudenbush and Bryk 2002, chapter 3). The p-value is less than 0.001 to reject the null hypothesis. Therefore, the CC analysis shows that the effects of the highest parent education vary randomly across schools, and that attending a private school, age, the highest parental occupation status, the highest parent education and family wealth are all positively associated with math achievement while student-to-teacher ratio and first-generation immigrant status are not significantly associated with the outcome.

4.2 Efficient Analysis

Now, we reanalyze the random coefficients model (34) given incomplete data by *HLM 7* according to the efficient missing data method explained in section 3. The ML estimates based on $m = 5$ imputations are displayed under the heading "Efficient" in Table 2. The Efficient analysis considered 5550 students attending 166 schools after dropping 61 students with the highest parent education missing because the method requires that the covariate having a random coefficient be completely observed. The τ_{33} estimate is 19 with the associated variance estimate 5.38^2, not shown in Table 2, that is less than the CC counterpart 6.45^2 above. Consequently, an approximate 95% confidence interval for τ_{33} is (10.93, 33.11) narrower and farther away from zero than the corresponding CC interval $(9.70, 36.90)$. To test $H_0 : \tau_{33} = 0$ against $H_a : \tau_{33} > 0$, we first note that the null hypothesis $\tau_{33} = 0$ implies $\tau_{03} = 0$ so that $k = 2$. *HLM 7* provides multiply imputed data sets. Given the tth completed data set based on the full model, we fit both the full and reduced models to obtain the likelihood ratio test statistic d_t and the average $\bar{d}$ to compute D_1 in Equation (25). The average relative increase in variance due to missing data in $Q = [\tau_{03}\ \tau_{33}]^T$ is $r_1 = 0.003 \approx 0$ to yield the test statistic $D_1 \approx \bar{d}/2$. Consequently, $2D_1 \approx 50.75$ gives the p-value $P(\chi^2_2 > 50.75) < 0.00001$ based on Equation (29). This method provides the range of the p-value between one half and twice the computed one (Li et al. 1991). This precision gives enough evidence to reject the null hypothesis in support of the alternative hypothesis that the effects of the highest parent education vary randomly across schools.

Parameter estimates with the associated standard errors, degrees of freedom and p-values are shown under "Efficient" in Table 2. The Efficient analysis shows that attending a private school, age, the highest parental occupation status, the highest parent education and family wealth are all positively associated with mathematics achievement as the CC analysis revealed. Controlling for the individual and school characteristics, however, the first-generation immigrant status is negatively associated with the outcome while the association is not statistically significant according to the CC analysis. A main reason for the different inferences is the lower standard error 4.61 of the Efficient analysis than the CC counterpart 5.27. Based on $m = 5$ imputations, the relative increase in variance due to missing data in the effect estimate is $r = (\sqrt{1191/4} - 1)^{-1} = 0.06$ based on the Equation (23). The fraction of missing information s in Equation (24) is lower than $r = 0.06$ so that the standard error 4.61 is at most $\sqrt{1 + 0.06/5} = 1.006$ times as high as the ideal one based on infinitely many imputations (Rubin 1987, p.114; Schafer 1997, p.107). Consequently, the effect estimate based on $m = 5$ imputations loses little precision, relative to the counterpart based on infinitely many imputations. That is, five imputations provide enough precision for estimation of the effect. Overall, the standard errors associated with the effect estimates of level-1 covariates under the Efficient analysis are up to 14% lower than the CC counterparts. In addition, the effect estimates 0.90 and 9.39 of age and family wealth under the Efficient analysis are considerably higher than their CC counterparts 0.64 and 6.48, respectively. Furthermore, the CC analysis seems to exaggerate the goodness of fit by producing smaller variance estimates than those of the Efficient analysis. At level 2, the CC analysis produces a lower standard error associated with the effect estimate of the private school indicator than does the Efficient analysis. The relatively understated CC standard error reflects its positive association with the comparatively underestimated CC variance components.

Based on the Efficient analysis, a typical non-immigrant student attending a public school with average age, highest parental occupation status, highest parent education, and family wealth scores 471 points in mathematics achievement on average. Students attending a private school score 34.51 points higher than do those attending a public school on average, controlling for the effects of other covariates in the model. One month older in age, a unit increase in the highest parental occupation status, one year increment in the highest parent education and a unit increase in family wealth are expected to raise mathematics scores by 0.90, 0.86, 4.18, and 9.39 points, respectively, ceteris paribus. Controlling for the individual and school characteristics in the model (34), the average difference in mathematics achievement between first-generation immigrants and other students reduces to 10.81 points or 30% of the initial gap, 36.08 points. Consequently, the individual and school characteristics considered in the model (34) explain 70% of the initial gap in mathematics achievement between first-generation immigrants and other students.

5 Data Analysis with Plausible Values

The Efficient analysis of HLM2 (34) in the previous section considers the first set of plausible mathematics scores as if they were completely observed. Consequently, the $m(=5)$ imputed or completed data sets reflect uncertainty due to the missing values of covariates, but does not take into account uncertainty from missing outcome values to produce understated standard errors of estimates. This section illustrates an efficient analysis strategy for the model (34) using all five sets of plausible mathematics scores. The first set has mean (standard deviation) equal to 475.18 (89.87) shown in Table 1. The second to the fifth sets have means (standard deviations) equal to 474.44 (89.09), 474.46 (88.95), 474.97 (88.66), and 474.54 (89.41), respectively.

In the previous section, we produced m imputations to efficiently analyze the desired HLM2 (34) with covariates subject to missingness given the first set of plausible outcome values. In this section, we repeat the same analysis with identical covariates given each set of plausible outcome values. With the number of q sets of plausible outcome values fixed at 5, this strategy produces $5m(= q \times m)$ completed data sets. Unlike the efficient analysis of the previous section based on the first set of plausible outcome values, the $5m$ imputations reflect uncertainty in parameter estimates due to missing values of both outcome and covariates. Note that we may obtain more imputations by increasing the number of m imputations per set of plausible outcome values. It is important flexibility to be able to increase m that will decrease the relative increase in variance due to missing data in Equation (20) and, thus, increase the degrees of freedom of an estimate in Equation (19), in particular, when the missing values of covariates account for a considerable amount of uncertainty in estimation. The degrees of freedom of an estimate is negatively associated with the p-value. This flexibility is not available to us for the outcome variable because the number of q sets of plausible outcome values is fixed at 5 by the imputer. Because the efficient analysis in the previous section reveals that uncertainty in estimation due to the missing values of covariates is not substantial, we generate $m = 1$ imputation per set of plausible outcome values to analyze 5 imputations in this section. Then we use the "Multiple Imputation" option of *HLM 7* that automates the complete-data analysis of the hierarchical model (34) given the multiple imputation to produce the combined estimates (Rubin 1987).

The ML estimates are displayed under the heading "Efficient PV" in Table 2. We compare the Efficient PV analysis with the Efficient analysis based on the first set of plausible values in the previous section. Again based on 5550 students attending 166 schools, the estimated τ_{33} is 14, lower than 19 produced under the Efficient analysis. The associated variance estimate is 6.09^2, higher than 5.38^2 under the Efficient analysis, to reflect added uncertainty due to missing outcome values. An approximate 95% confidence interval for τ_{33} is (5.97,32.85) wider and closer to zero than the Efficient analysis counterpart (10.93, 33.11). For testing $H_0 : \tau_{33} = 0$ versus $H_a : \tau_{33} > 0$, the average relative increase in variance due to missing data in $Q = [\tau_{03}\ \tau_{33}]^T$ is $r_1 = 0.61$ based on Equation (26) to yield the test statistic $D_1 = 13.73$. Higher than the corresponding $r_1 = 0.003$ based on the first set of plausible values in the previous section, the $r_1 = 0.61$ implies that missing outcome values add a considerable amount of uncertainty to the Q estimates. The p-value (28) is $P(F_{2,18} > 13.73) = 0.0002$ to reject the null hypothesis in support of the alternative hypothesis that $\tau_{33} > 0$.

Both Efficient PV analysis and Efficient analysis produce comparable effect estimates and the same statistical inferences. However, the Efficient PV analysis yields comparatively low degrees of freedom overall to reveal added uncertainty in the estimates due to the missing values of the outcome variable. In particular, the degrees of freedom for the effect estimate of the first-generation immigrant status reduce from 1191 under the Efficient analysis to 66 under the Efficient PV analysis. The 66 degrees of freedom translate into the relative increase in variance due to missing data $r = \left(\sqrt{66/4} - 1\right)^{-1} \approx 0.33$ based on Equation (23), which is a substantial increase from the corresponding $r = 0.06$ under the Efficient analysis. The relative increase in r implies that added uncertainty due to missing outcome values is considerable, thereby inflating the standard error 4.61 and the p-value 0.019 under the Efficient analysis to 5.08 and 0.045 under the Efficient PV analysis, respectively.

6 Extensions and Limitations

This chapter explained how to efficiently analyze a two-level hierarchical linear model where explanatory as well as outcome variables may be subject to missingness with a general missing pattern at any of the levels (Shin and Raudenbush 2007). The key idea is to reexpress the desired hierarchical model as the joint distribution of variables, including the outcome, that are subject to missingness conditional on all of the covariates that are completely observed under multivariate normality; estimate the joint distribution by ML; generate multiple imputation given the ML estimates of the joint distribution; analyze the desired hierarchical model by complete data analysis given the multiple imputation; and then combine the multiple hierarchical model estimates (Rubin 1987). Given the desired hierarchical model specified by a data analyst, the rest of the analysis steps can be automated for efficient estimation of the hierarchical model. The automation has been implemented in a software package *HLM 7* that is yet to be released to the public at the time of writing this chapter. With such a software package in hand, multilevel incomplete data analysis is no different from complete data analysis from the data analyst's perspective.

This chapter illustrated two examples for efficient analysis of incompletely observed PISA data with *HLM 7*. The outcome variable was a mathematics achievement score subject to missingness. PISA imputed five sets of plausible values for each missing outcome value. Assuming the first set of plausible mathematics scores completely observed in the first example, we efficiently analyzed hierarchical linear model (34) given covariates subject to missingness at multiple levels. We compared the efficient analysis with the complete-case analysis. Overall, the complete-case analysis produced higher standard errors than did the efficient analysis, and some estimates of the complete-case analysis were considerably different from the counterparts of the efficient analysis. Consequently, the two analyses produced different statistical inferences for the effect of a key covariate. The second example was analysis of the same model (34) with covariates subject to missingness and all plausible outcome values. We repeated the efficient analysis of the first example with identical covariates given each set of plausible outcome values, each time imputing a single completed data set to eventually produce as many completed data sets as the five sets of plausible outcome values for subsequent complete-data analysis. The combined estimates were comparable to and produced the same statistical inferences as those of the efficient analysis in the first example. On the other hand, the degrees of freedom for estimates were considerably lower than those of the first example, overall. That is, the estimates exhibited higher relative increase in variance due to missing data than did those of the first example based on the first set of plausible values. Consequently, a substantial amount of uncertainty due to missing data in estimation was from missing outcome values.

The second example illustrates one of the difficulties in incomplete data analysis when the imputer of the plausible outcome values is different from the data analyst of the desired hierarchical model (34) (Meng 1994; Rubin 1996). With $q = 5$ sets of plausible outcome values and covariates subject to missingness at multiple levels, the analysis based on 5 imputations ($m = 1$ imputation per set of plausible outcome values) reveals that a considerable amount of uncertainty due to missing data in estimation is from missing outcome values. Consequently, the data analyst may want to increase the number of q imputations of plausible outcome values, which may increase the degrees of freedom and reduce the computed p-value because the degrees of freedom (19) of an estimate is positively associated with the number of imputations. However, with access to 5 sets of plausible outcome values only, she is unable to reduce the relative increase in variance due to missing outcome values. Increasing the number of m imputations per set of plausible outcome values may help reduce the relative increase in variance due to the missing values of covariates, not due to missing outcome values. When both outcome and covariates contain missing values, the efficient missing data method and its software introduced in this chapter provide flexibility to manipulate the number of imputations. In that case, the data analyst is able to set the number of imputations and produce multiple imputation tailored to the analyst's own analysis with the software. Because the analyst is also the imputer, the analyst will not face the difficulty that arises when he or she is not the imputer (Meng 1994; Rubin 1996).

The two-level missing data approach in this chapter has been extended to efficient analysis of a two-level contextual-effects model (Raudenbush and Bryk 2002; Shin and Raudenbush 2010); of a three-level hierarchical linear model (Shin and Raudenbush 2011; Shin 2012) and of an arbitrary Q-level hierarchical linear model (Shin and Raudenbush 2013) given incomplete data. Three-level user-friendly software based on the missing data methods of Shin and Raudenbush (2011) and Shin (2012) is under development at the time of writing this chapter. These advances guide us with continuous variables subject to missingness.

The efficient analysis of HLM2 (34) in Table 2 involves discrete first-generation immigrant status and private school indicator subject to missingness at levels 1 and 2, respectively. Although it is not appropriate to handle the discrete missing data under the corresponding multivariate joint normal distribution (13) of variables subject to missingness, the implied conditional model is the desired hierarchical model (34). Furthermore, the joint model assumption (13) to handle missing data affects only imputed data, not observed data. The advantage is that the hierarchical model is analyzed by the efficient missing data method (Schafer 1997; Shin and Raudenbush 2007). However, with the missing rate high, the normal joint model assumption becomes nontrivial. A useful future extension of this approach is to efficient and robust handling of normal and nonnormal multilevel missing data.

It entails challenges to extend the missing data method introduced in this chapter to analysis of multilevel incomplete data from multistage sampling where different selection probabilities of units are used (OECD 2007; Tourangeau el al. 2009). The extent to which complicated sampling weights affect the missing data analysis is not well known. To minimize possible adverse impact such as biased inferences, the sampling weights may be applied at the final stage of complete-data analysis given multiple imputation (Graubard and Korn 1996; Pfeffermann et al. 1998; Korn and Graubard 2003). An important future research topic is extension of the efficient missing data method to analysis of multilevel incomplete data generated from multistage sampling with different selection probabilities of units.

Another limitation of the missing data method is that the covariates having random coefficients must be completely observed. With such covariates subject to missingness, the joint distribution of variables subject to missingness may not be expressed as a multivariate normal distribution. Consequently, subsequent analysis given the estimated normal joint model by ML does not apply. Research is under way to relax the assumption, which will broaden the applicability of this method.

Acknowledgements This work 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 author and do not represent views of the Institute or the U.S. Department of Education.

References

[1] Asparouhov, T. and Muthén, B. (2010). Multiple Imputation with Mplus. Technical Report. www.statmodel.com.

Table 1: US data for analysis from PISA data collection in 2006.

Level	Variables	Description	mean (SD[a])	missing (%)
I	MATH	mathematics scores	475.18 (89.87)	0 (0)
	AGE	age in months	189.79 (3.54)	0 (0)
	HISEI	highest parental occupation status	52.46 (16.78)	390 (7)
	PARENDED	highest education level of parents in the number of years of schooling	13.61 (2.49)	61 (1)
	WEALTH	family wealth	0.15 (0.80)	34 (1)
	IMMIG1	1 if 1st-generation immigrant	0.06 (0.24)	189 (3)
II	STRATIO	student-to-teacher ratio	15.46 (4.64)	28 (17)
	PRIVATE	1 if private	0.09 (0.29)	3 (2)

[a] standard deviation

[2] Browne, W.J. (2012). MCMC estimation in MLwiN Version 2.25. Centre for Multilevel Modelling, University of Bristol.

[3] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM Algorithm. *Journal of the Royal Statistical Society B*, **76**, 1-38.

[4] Dempster, A.P., Rubin, D.B. and Tsutakawa, R.K. (1981). Estimation in Covariance Components Models. *Journal of the American Statistical Association*, **76**, 341-353.

[5] Diggle, P. and Kenward, M.G. (1994). Informative Drop-Out in Longitudinal Data Analysis. *Journal of the Royal Statistical Society C*, **43**, 49-93.

[6] Goldstein, H. and Browne, W. (2002). Multilevel factor analysis modelling using Markov Chain Monte Carlo estimation. In Marcoulides, G. and Moustaki, I. eds. *Latent variable and latent structure models*, London: Lawrence Erlbaum.

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

Table 2: Analysis of the random-coefficients model (34).

Covariate	CC[a]	Efficient[b]			Efficient PV[c]		
	coef. (se)[d]	coef.(se)	df[e]	p-value	coef.(se)	df	p-value
Intercept	478 (3.15)[f]	471 (3.11)[f]	163	< 0.001	471 (3.14)[f]	163	< 0.001
STRATIO	-0.11 (0.60)	-0.18 (0.68)	99	0.792	-0.24 (0.70)	49	0.729
PRIVATE	35.85 (10.19)[f]	34.51 (11.36)[f]	163	0.003	36.52 (11.44)[f]	163	0.002
AGE	0.64 (0.32)[f]	0.90 (0.29)[f]	5214	0.002	0.97 (0.30)[f]	494	0.001
HISEI	0.82 (0.08)[f]	0.86 (0.08)[f]	124	< 0.001	0.84 (0.09)[f]	26	< 0.001
PARENTED	4.93 (0.70)[f]	4.18 (0.61)[f]	165	< 0.001	4.24 (0.66)[f]	61	< 0.001
WEALTH	6.48 (1.62)[f]	9.39 (1.42)[f]	5214	< 0.001	9.25 (1.43)[f]	2058	< 0.001
IMMIG1	-10.04 (5.27)	-10.81 (4.61)[f]	1191	0.019	-10.37 (5.08)[f]	66	0.045
τ	$\begin{bmatrix} 1027 & 103 \\ 103 & 19 \end{bmatrix}$	$\begin{bmatrix} 1280 & 117 \\ 117 & 19 \end{bmatrix}$			$\begin{bmatrix} 1288 & 111 \\ 111 & 14 \end{bmatrix}$		
σ^2	5229	5360			5252		

[a] Complete-case analysis.
[b] Efficient analysis of the first set of plausible outcome values.
[c] Efficient analysis of all plausible outcome values.
[d] Coefficient (standard error).
[e] Degrees of freedom.
[f] Statistically significant at the significance level $\alpha = 0.05$.

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

[9] Graubard, B.I., and Korn, E.L. (1996). Modelling the sampling design in the analysis of health surveys. *Statistical Methods in Medical Research*, **5**, 263-281.

[10] Horton, N.J. and Kleinman, K.P. (2007). Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models. *The American Statistician*, **61**, 79-90.

[11] Korn, E.L., and Graubard, B.I. (2003). Estimating variance components by using survey data. *Journal of the Royal Statistical Society B*, **65**, 175-190.

[12] Krueger, A.B. (1999). Experimental Estimates of Education Production Functions. *The Quarterly Journal of Economics*, **114**, 497-532.

[13] Li, K., Meng, X., Raghunathan, T.E. and Rubin D.B. (1991). Significance Levels from Repeated p-Values with Multiply-Imputed Data. *Statistica Sinica*, **1**, 65-92.

[14] Little, R.J.A. (1995). Modeling the Drop-Out Mechanism in Repeated-Measures Studies, *Journal of the American Statistical Association*, **90**, 1112-21

[15] Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, New York: Wiley.

[16] Little, R. J. A. (2009). Selection and pattern-mixture models. In G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs (Eds.), Longitudinal data analysis (pp. 409-431). Boca Raton, FL: Chapman & Hall/CRC.

[17] Lunn, D., Spiegelhalter, D., Thomas, A., and Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, **28**, 3049-82.

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

[19] Meng, X. and Rubin, D.B. (1992). Performing Likelihood Ratio Tests with Multiply-Imputed Data Sets. *Biometrika*, **79**, 103-111.

[20] Meng, X. (1994). Multiple-Imputation Inferences with Uncongenial Sources of Input. *Statistical Science*, **9**, 538-558.

[21] Muthén, L. and Muthén, B. (2010). Mplus users guide (6th ed.). Los Angeles, CA: Muthén and Muthén.

[22] Muthén, B., Asparouhov, T., Hunter, A.M. and Leuchter, A.F. (2011). Growth Modeling With Nonignorable Dropout: Alternative Analyses of the STAR*D Antidepressant Trial. *Psychological Methods*, **16**, 17-33.

[23] OECD (2007), PISA 2006: Science Competencies for Tomorrow's Wrold, OECD, Paris.

[24] Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H. and Rasbash, J. (1998). Weighting for Unequal Selection Probabilities in Multilevel Models. *Journal of the Royal Statistical Society B*, **60**, 23-40.

[25] Pinheiro, J.C. and Bates, D.M. (2000). *Mixed-Effects Models in S and S-plus*, New York: Springer.

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

[27] Rasbash, J., Steele, F., Browne, W. and Goldstein, H. (2009). A Users Guide to MLwiN Version 2.10, Centre for Multilevel Modelling, University of Bristol.

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

[29] Raudenbush, S.W., Bryk, A.S., Cheong, Y., Congdon, R.T. and Mathilda du Toit (2011). *HLM 7: Hierarchical Linear and nonlinear modeling.* Lincolnwood, IL: Scientific Software International.

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

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

[32] Rubin, D.B. (1996). Multiple Imputation After 18+ Years. *JASA*, **91**, 473-489.

[33] Rubin, D.B. (2003). Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica*, **57**, 3-18.

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

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

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

[37] Schafer, J. L. (2003). Multiple imputation in multivariate problems when imputation and analysis models differ. *Statistica Neerlandica*, **57**, 19-35.

[38] Shin, Y. (2003). *Inference and Applications in Hierarchical Linear Models with Missing Data.* Unpublished doctoral dissertation, University of Michigan.

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

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

[41] Shin, Y. and Raudenbush, S. W. (2011). The Causal Effect of Class Size on Academic Achievement: Multivariate Instrumental Variable Estimators with Data Missing at Random. *Journal of Educational and Behavioral Statistics*, **36**, 154-185.

[42] Shin, Y. (2012). Do Black Children Benefit More from Small Classes? Multivariate Instrumental Variable Estimators with Ignorable Missing Data. *Journal of Educational and Behavioral Statistics*, **37**, 543-574.

[43] Shin, Y. and Raudenbush, S. W. (2013). Efficient Analysis of Q-Level Nested Hierarchical General Linear Models Given Ignorable Missing Data. To appear in *International Journal of Biostatistics*.

[44] Snijders, T.A.B. and Bosker, R.J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, London: Sage.

[45] Spiegelhalter, D.J., Thomas, A. and Best, N.G. (2000). WinBUGS version 1.3: user manual. Medical Research Council Biostatistics Unit, Cambridge.

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

[47] Tourangeau, K., Nord, C., Lê, T., Sorongon, A.G., and Najarian, M. (2009). *Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), Combined User's Manual for the ECLS-K Eighth-Grade and K-8 Full Sample Data Files and Electronic Codebooks* (NCES 2009-004). NCES, IES, DOE. Washington, D.C.

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

[49] Van Buuren, S., and Groothuis-Oudshoorn, K. (2011). MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, **45**, 1-67.

[50] Van Buuren, S. (2011). Multiple imputation of multilevel data. In Hox, J.J. and Roberts, J.K. (Eds.), The Handbook of Advanced Multilevel Analysis, Chapter 10, pp. 173-196. Milton Park, UK: Routledge.

[51] Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*, New York: Springer.

[52] Yuan, Y. C. (2000). Multiple imputation for missing data: Concepts and new development. In *Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference* (Paper No. 267). Cary, NC: SAS Institute.

[53] Yucel, R. M. (2008). Multiple imputation inference for multivariate multilevel continuous data with ignorable non-response. Philosophical Transactions of the Royal Society A, 366, 2389-2403.