CHAPTER 16

MISSING DATA IN MULTILEVEL RESEARCH

Simon Grund, Oliver Lüdtke, and Alexander Robitzsch

Multilevel data are often incomplete, for example, when participants refuse to answer some items in a questionnaire or drop out of a study that involves multiple measurement occasions. Even though there is a consensus that current state-of-the-art procedures for statistical analyses with missing data should be preferred (e.g., Allison, 2001; Enders, 2010; Little & Rubin, 2002; Newman, 2014; Schafer & Graham, 2002), simpler methods such as listwise deletion (LD) prevail and are still widely applied in research practice (Jeličić, Phelps, & Lerner, 2009; Nicholson, Deboeck, & Howard, 2017; Peugh & Enders, 2004). This is problematic because these methods can distort parameter estimates and statistical inference. In this chapter, we provide a general introduction to the problem of missing data in multilevel research, and we present two principled methods for handling incomplete data: multiple imputation (MI) and maximum likelihood (ML) estimation. We discuss how these procedures can be used to address missing data in multilevel research, and we consider their commonalities as well as their individual strengths and weaknesses. A brief computer simulation study is used to illustrate the statistical behavior of the parameter estimates obtained from these methods. Finally, we illustrate their application in a data analysis example and provide the syntax files and computer code needed to reproduce our results.

EXAMPLE: JOB SATISFACTION AND LEADERSHIP STYLE

To provide an illustration of the ideas presented here, we adopt a running example in which we examine the relationships between job satisfaction and several work-related variables. For the purpose of this chapter, we regard the multilevel structure as cross-sectional, for example, with employees at Level 1 nested within work groups at Level 2. The example is based on the data from Klein et al. (2000). The study features a sample of 750 employees from 50 work groups with measures of job satisfaction (*SAT*), negative leadership style (*LS*), workload (*WL*), and cohesion (*COH*). We altered the data set slightly by (a) transforming workload into a categorical variable (high vs. low) and (b) treating cohesion as a *global* variable that was directly assessed at Level 2 (e.g., a supervisor rating). We investigated the relationships between employees' job satisfaction and negative leadership style, workload, and cohesion using a multilevel random intercept model (Snijders & Bosker, 2012). In the hierarchical notation of Raudenbush and Bryk (2002), the Level 1 equation of the model reads

$$SAT_{ij} = \beta_{0j} + \beta_{1j}\left(LS_{ij} - \overline{LS}_{\cdot j}\right) + \beta_{2j}WL_{ij} + r_{ij} \tag{16.1}$$

http://dx.doi.org/10.1037/0000115-017
The Handbook of Multilevel Theory, Measurement, and Analysis, S. E. Humphrey and J. M. LeBreton (Editors-in-Chief)

with Level 2 equations

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + \gamma_{01}\overline{LS}_{\cdot j} + \gamma_{02}COH_j + u_{0j}\\ \beta_{1j} &= \gamma_{10}\\ \beta_{2j} &= \gamma_{20}.\end{aligned} \tag{16.2}$$

Here, SAT_{ij} denotes the job satisfaction of an employee i in group j. The ratings on leadership style were subjected to group-mean centering, where LS_{ij} denotes employees' individual ratings of leadership style, and $\overline{LS}_{\cdot j}$ denotes the average rating in group j. Finally, WL_{ij} denotes employees' workload, and COH_j denotes a work group's cohesion (e.g., a supervisor rating). The random intercept, u_{0j} and the residuals, r_{ij} were each assumed to follow a normal distribution with mean zero and variances τ_0^2 and σ^2, respectively. In the remainder of this chapter, we will express this model with a combined notation (e.g., Snijders & Bosker, 2012)

$$\begin{aligned}SAT_{ij} = \gamma_{00} + \gamma_{10}\left(LS_{ij} - \overline{LS}_{\cdot j}\right) + \gamma_{01}\overline{LS}_{\cdot j}\\ + \gamma_{20}WL_{ij} + \gamma_{02}COH_j + u_{0j} + r_{ij}.\end{aligned} \tag{16.3}$$

In this chapter, we focus on multilevel models in which only the intercept varies across groups. Longitudinal research designs as well as multilevel models with additional random effects (e.g., random slopes) are considered in the Discussion section.

MISSING DATA IN MULTILEVEL RESEARCH

It is well known that simpler methods of dealing with missing data (e.g., LD) can severely compromise statistical decision making (e.g., Enders, 2010; Little & Rubin, 2002). For example, when analyses are based on only the complete cases, then parameter estimates can be biased (i.e., the estimates may systematically differ from the "true" values that hold in the population) when data are missing in a systematic manner (e.g., see Schafer & Graham, 2002). However, even when data are missing in an unsystematic manner, inferences based on LD are often inefficient (i.e., low statistical power) due to the reduction in sample size and because potentially useful information about the missing data is being ignored (e.g., Newman, 2014). Therefore, the common goals of the "principled" methods for handling missing data (e.g., ML and MI) are to (a) provide unbiased estimates for the statistical parameters of interest, (b) acknowledge the uncertainty that is due to missing data, and (c) make full use of the data in order to limit the loss of efficiency. However, before we devote ourselves to explaining these methods, it will be useful to first establish a formal framework for discussing the missing data problems and the challenges that can arise in multilevel research. In the following section, we discuss (a) possible mechanisms that can lead to missing data and (b) different patterns of missing data that can occur in multilevel data.

Missing Data Mechanisms

Rubin (1976) considered three broad classes of missing data mechanisms. We assume that there is a hypothetical complete data set, **Y**, which can be decomposed into an observed part, $\mathbf{Y}_{obs}$, and an unobserved part, $\mathbf{Y}_{mis}$, where an indicator matrix, **R**, denotes which elements are observed and which ones are missing. Rubin defined data to be missing at random (MAR) when the probability of observing data, $P(\mathbf{R})$, is independent of the missing data given the observed data, that is, $P(\mathbf{R}|\mathbf{Y}) = P(\mathbf{R}|\mathbf{Y}_{obs})$. In other words, under MAR, no link remains between the chance of observing data and the data themselves (i.e., they occur at random) once the observed data are taken into account. A special case of this scenario occurs when the probability of missing data is completely independent of the data, that is, $P(\mathbf{R}|\mathbf{Y}) = P(\mathbf{R})$, which is referred to as missing *completely* at random (MCAR). By contrast, when the probability of missing data is related to the unobserved data, that is, $P(\mathbf{R}|\mathbf{Y}) = P(\mathbf{R}|\mathbf{Y}_{obs},\mathbf{Y}_{mis})$, it is more difficult to infer from incomplete data and strong assumptions must be made about the missing data mechanism (see Carpenter & Kenward, 2013; Enders, 2011). This is referred to as missing *not* at random (MNAR).

The meaning of these mechanisms can be subtle, and they are best explained in an example (see also Enders, 2010). Consider the simple scenario illustrated in Figure 16.1, where negative leadership style is associated with lower job satisfaction, and

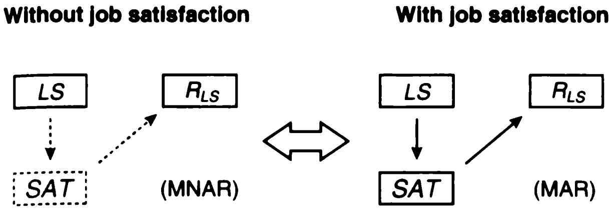

FIGURE 16.1. Example of systematic data loss and the effects of ignoring possible causes of missing data. LS = leadership style; MAR = missing at random; MNAR = missing not at random; R_{LS} = indicator for missing values in leadership style; SAT = job satisfaction.

ratings on leadership style are missing (R_{LS}) as a function of job satisfaction, say, because employees with low job satisfaction were less willing to answer questions about their supervisors (single-headed arrows). In this scenario, larger values of leadership style would be more likely to be missing (double-headed arrow), rendering statements about this variable misleading as long as they do not take the missing data mechanism into account (left panel). For example, the estimated mean of leadership style may be well below the "true" mean because larger values have a higher chance of being missing. However, with job satisfaction taken into account, these ties are broken (right panel): Given the values of job satisfaction, the scores of leadership style are now MAR, allowing us to estimate the *conditional* mean of leadership style given job satisfaction (e.g., using linear regression) and to make statements about the overall mean on this basis (see also Carpenter & Kenward, 2013).

The notion of missing data mechanisms allows us to identify conditions under which a missing data treatment may yield more or less accurate results in some model of interest. For example, LD generally provides unbiased estimates for a model of interest only under MCAR (see also Newman, 2014). In addition, LD may provide unbiased results in some very specific scenarios in which data are MAR or MNAR (e.g., Galati & Seaton, 2016; Little, 1992). However, because the assertion of specific missing data mechanisms requires untestable assumptions to be made, LD should be avoided in favor of procedures that make full use of the data and that are applicable under a more general set of assumptions (e.g., ML and MI; see also Schafer & Graham, 2002). Both ML and MI provide unbiased results under MAR. In such a case, the exact mechanism need not be known and may even be different from individual to individual as long as the observed data are sufficient to "break the link" between the unobserved data and the probability that they are missing (Carpenter & Kenward, 2013). To make this assumption more plausible, it is often recommended that auxiliary variables be included in the treatment of missing data. Such variables are not part of the model of interest but are related to the probability of missing data or the variables with missing data themselves (Collins, Schafer, & Kam, 2001; Enders, 2008; Graham, 2003). Including such variables is beneficial because (a) they make the MAR assumption more plausible, and (b) if they are related to the variables of interest, they provide information about the missing values and improve statistical power (Collins et al., 2001).

Patterns of Missing Data

For the treatment of missing data, it can also be useful to distinguish different *patterns* of missing data. Such a distinction may help researchers to identify problems with the data and navigate choices regarding the missing data treatment. In accordance with Newman (2014), we distinguish three basic patterns: *item*, *construct*, and *unit* nonresponse. Item nonresponse denotes cases in which participants fail to answer a single item on a questionnaire (e.g., an item concerning salary from a questionnaire for assessing job satisfaction). By contrast, construct and unit nonresponse, respectively, denote cases in which all items pertaining to a certain construct or even a participant's entire questionnaire may be missing (e.g., because a participant was absent on the day the company conducted a survey). In the present chapter, we focus on item nonresponse, although construct nonresponse can often be addressed by applying similar methods (see also Gottschall, West, & Enders, 2012). Unit nonresponse can be more complicated to deal with and is often addressed by employing survey weights (e.g., Särndal, Swensson, & Wretman, 2003).

In multilevel research, item, construct, and unit nonresponse can occur at different levels of the sample (see also van Buuren, 2011). According to Kozlowski and Klein (2000), we may again distinguish three different patterns of missing data. Data can be missing (a) at Level 1, (b) in *global* variables at Level 2, or (c) in *shared* variables at Level 2. Missing data at Level 1 refer to the lowest level of the sample (e.g., missing data from employees). Global variables refer to variables that are directly assessed at Level 2 (e.g., missing data in supervisor rating), whereas shared variables denote variables that are assessed at Level 1 and then aggregated at Level 2 (e.g., a group average based on incomplete data collected from employees). Because missing data both at Level 1 and in shared variables at Level 2 originate at Level 1, they can usually be addressed by the same methods. Missing data in global variables sometimes require additional considerations but can be treated with similar tools. Additional patterns of missing data are possible (e.g., incomplete data about group membership), but these will not be our focus in the present chapter (Goldstein, 2011; for a discussion, see Hill & Goldstein, 1998).

For example, consider Table 16.1. In the first group of employees, only a single response to the workload variable is missing (Level 1, item missing). In the second group, the ratings on leadership style are missing for all employees (Level l, item missing), and the group mean is missing as a result (shared Level 2, item missing). In the third group, one employee did not respond to any items (Level 1, unit missing). In that group, the group mean might be calculated from the observed values, but it will be subject to uncertainty and possible bias because the underlying items are incomplete (shared Level 2, item missing). In addition, the cohesion score is missing for all employees in that group (global Level 2, item missing). Finally, the last employee could not be assigned to a group with sufficient certainty.

TABLE 16.1

Hypothetical Example of a Pattern of Missing Data in a Multilevel Sample

Case	Group	SAT_{ij}	LS_{ij}	WL_{ij}	COH_j	$\overline{LS}_{.j}$
1	1	2.3	?	High	3.8	?
2	1	1.7	?	Low	3.8	?
3	1	1.7	?	High	3.8	?
4	2	1.8	2.3	Low	?	2.2
5	2	1.4	2.1	High	?	2.2
6	2	?	?	?	?	2.2
7	3	3.4	1.2	Low	2.7	1.4
8	3	2.8	1.8	?	2.7	1.4
9	3	3.1	1.2	Low	2.7	1.4
10	?	2.1	2.3	High	?	?

Note. Missing observations are indicated by question marks. *COH* = cohesion; *LS* = leadership style; *SAT* = job satisfaction; *WL* = workload.

METHODS FOR HANDLING MISSING DATA

In this section, we consider two general procedures that are currently regarded as principled methods for handling missing data (e.g., Schafer & Graham, 2002). First, we consider MI. We elaborate on different approaches to multilevel MI, and we discuss potential challenges when specifying imputation models for multilevel data. As a second procedure, we consider the estimation of multilevel models by ML. Finally, we provide a comparison of the two procedures from a practical point of view.

Multiple Imputation

The basic idea in MI is to replace missing values with an "informed guess" obtained from the observed data and a statistical model (the imputation model). A schematic representation of this process is displayed in Figure 16.2. Multiple imputation generates several (M) replacements for the missing data by drawing from a predictive distribution of the missing data, given the observed data and the parameters of the imputation model. The M data sets are then analyzed separately, yielding M sets of parameter estimates (i.e., $\hat{Q}_1, \ldots, \hat{Q}_M$), and these are combined into a set of final parameter estimates (i.e., $\hat{Q}_{MI}$) and inferences using the rules outlined by Rubin (1987).

When performing MI, the imputation model must be chosen in such a way that it "matches" the model of interest, that is, it must be specified in such a way that it preserves the relationships among variables and the relevant features of the analysis model (Meng, 1994; Schafer, 2003). For example, if the model of interest is a regression model with

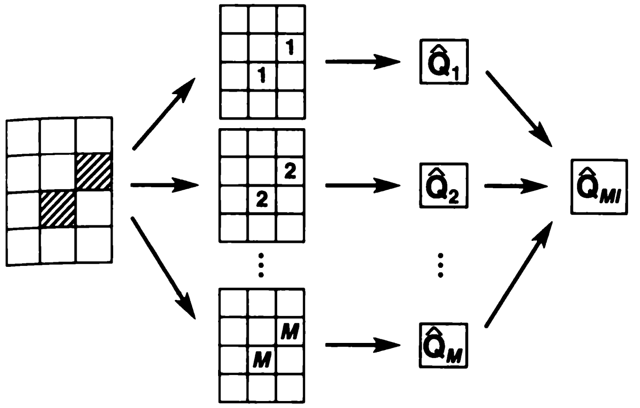

FIGURE 16.2. Schematic representation of multiple imputation (MI) and the analysis of multiply imputed data sets. $\hat{Q}$ = estimator of the parameter of interest.

an interaction effect, then the imputation model must also include the interaction; otherwise, it will be more difficult to detect the interaction effect in subsequent analyses (Enders, Baraldi, & Cham, 2014). In multilevel research, it is important for the imputation model to incorporate the multilevel structure of the data. In the following, we review different strategies for accommodating the multilevel structure during MI, including ad hoc strategies on the basis of single-level MI. We consider two broad approaches to MI: joint modeling and the fully conditional specification of MI. In the joint modeling approach, a single statistical model is specified for all incomplete variables simultaneously. In the fully conditional specification, each variable is imputed in turn using a sequence of models (for a discussion, see Carpenter & Kenward, 2013). Finally, we discuss strategies for analyzing multiply imputed data sets and pooling their results.

Strategies based on single-level multiple imputation. Perhaps the simplest approach to multilevel MI is to ignore the multilevel structure of the data and employ single-level MI. With this strategy, the multilevel structure is disregarded altogether. Not surprisingly, it has been shown that single-level MI can lead to biased estimates in subsequent multilevel analyses (Black, Harel, & McCoach, 2011; Enders, Mistler, & Keller, 2016; Taljaard, Donner, & Klar, 2008). Lüdtke, Robitzsch, and Grund (2017) demonstrated that single-level MI tends to underestimate the intraclass correlation (ICC; also known as the ICC(1)) of variables with missing data and may either under- or overestimate within- and between-group effects in multilevel random intercept models. Figure 16.3 shows the expected bias in the ICC of a variable Y relative to its true value (i.e., in percent) and for different numbers of individuals per group (n), different values of the ICC of Y and an auxiliary variable X, and different amounts of missing data (25%, 50%). As can be seen, single-level MI tends to underestimate the true ICC. For example, in the

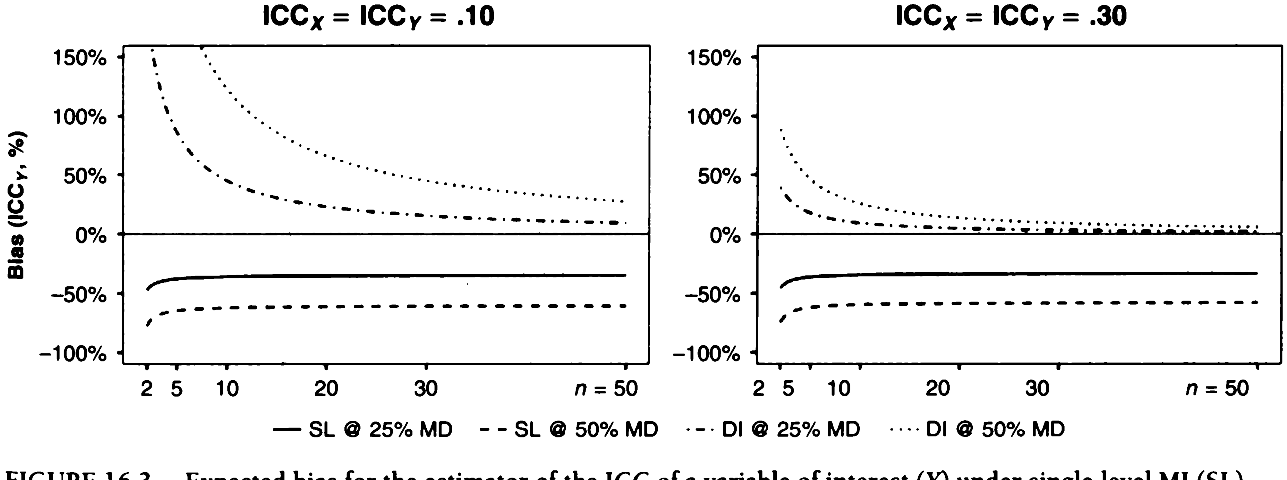

FIGURE 16.3. Expected bias for the estimator of the ICC of a variable of interest (Y) under single-level MI (SL) and the dummy-indicator approach (DI). It is assumed that all groups contain the same number of individuals (n) and the same proportion of missing data (MD) in Y. ICC_X = intraclass correlation of an auxiliary variable; ICC_Y = intraclass correlation of the variable of interest.

scenario with $n = 5$ individuals per group and 25% missing data, single-level MI is expected to yield an estimate of only .062 when the true ICC is .100 and of only .191 when the true ICC is .300. In either case, the true ICC is underestimated by approximately 37%.

To remedy this situation, it has been suggested that the multilevel structure be represented by a number of dummy indicator variables (i.e., the DI approach; e.g., Graham, 2009). This strategy effectively estimates a separate group mean for each group by estimating the imputation model conditional on group membership, thus incorporating group differences during MI (see also Enders et al., 2016). For example, the differences in job satisfaction between the 50 work groups in our running example can be represented in a regression model by the intercept and an additional 49 dummy variables (with one group selected as a reference group). The performance of this strategy depends on the situation in which it is applied. As demonstrated by Drechsler (2015), the DI approach tends to overestimate the ICC of variables with missing data but yields approximately unbiased estimates of the regression coefficients in a multilevel analysis model when missing data are restricted to the dependent variable (see also Andridge, 2011). However, because the DI approach exaggerates the variance between groups, it provides only a biased estimate of the between-group effect if missing values occur in explanatory variables (Lüdtke et al., 2017). As shown in Figure 16.3, the DI approach tends to overestimate the true ICC. The bias is particularly strong when the true ICC is small and there are only a few individuals per group. For example, with $n = 5$ individuals per group and 25% missing data, the DI strategy is expected to yield an estimate of around .186 when the true ICC is .100 and of around .353 when the true ICC is .300. This corresponds to overestimations of the ICC by approximately 86% and 18%, respectively.

Joint modeling. To accommodate the nested structure of multilevel data, it has been recommended that MI be performed by using mixed-effects models (e.g., Enders et al., 2016; Lüdtke et al., 2017; Yucel, 2008). In the joint modeling approach to multilevel MI, a single model is specified for all variables with and without missing data, and imputations are generated from this model for all variables simultaneously.[1] The joint model can be regarded as a multivariate extension of univariate multilevel models; that is, it addresses multiple dependent variables simultaneously. The model reads

$$\begin{aligned} \mathbf{y}_{1ij} &= \boldsymbol{\gamma}_1 + \mathbf{u}_{1j} + \mathbf{r}_{1ij} && \text{(Level 1)} \\ \mathbf{y}_{2j} &= \boldsymbol{\gamma}_2 + \mathbf{u}_{2j}, && \text{(Level 2)} \end{aligned} \qquad (16.4)$$

where $\mathbf{y}_{1ij}$ denotes a vector of responses for individual i in group j with fixed intercepts $\boldsymbol{\gamma}_1$, random intercepts $\mathbf{u}_{1j}$, and residuals $\mathbf{r}_{ij}$. Similarly, $\mathbf{y}_{2j}$ denotes a vector of responses for group j (i.e., global variables) with fixed intercepts $\boldsymbol{\gamma}_2$ and residuals $\mathbf{u}_{2j}$. The random effects and residuals at Level 2 ($\mathbf{u}_{1j}$,$\mathbf{u}_{2j}$) are assumed to jointly follow a multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Psi}$. The residuals at Level 2 follow a multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}$. The joint model was originally developed by Schafer and Yucel (2002) to treat missing data at Level 1 and has since been extended to address missing data in categorical variables and variables at Level 2 (Asparouhov & Muthén, 2010; Carpenter & Kenward, 2013; Goldstein, Carpenter, Kenward, & Levin, 2009).

To illustrate how the joint model accommodates the multilevel structure, consider our running example and the illustration in Figure 16.4. The model of interest is a random intercept model that includes variables assessed at Levels 1 and 2 as well as relations between job satisfaction and leadership style both within and between groups (Equation 16.3). The joint model includes all variables as dependent variables in a multivariate random intercept model (Figure 16.4). For each variable at Level 1, the model includes a random

[1] The joint model can be expressed in a more general way, which allows fully observed variables to be included as predictor variables on the right-hand side of the model. However, in the present chapter, we consider only the "empty" specification of the model because it is easy to specify and widely applicable in the context of multilevel random intercept models (Enders et al., 2016; for a discussion, see Grund, Lüdtke, & Robitzsch, 2016b).

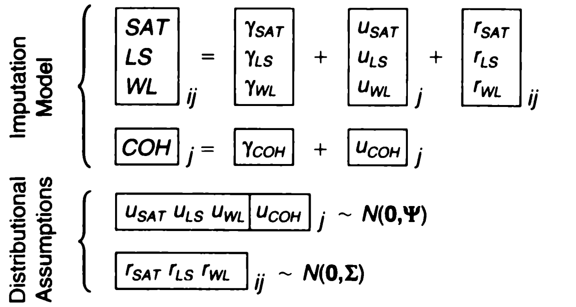
Imputation Model
Distributional Assumptions
SAT LS WL ij = γSAT γLS γWL + uSAT uLS uWL j + rSAT rLS rWL ij
COH j = γCOH + uCOH j
uSAT uLS uWL uCOH j ~ N(0,Ψ)
rSAT rLS rWL ij ~ N(0,Σ)

FIGURE 16.4. Schematic representation of the joint imputation model and its distributional assumptions in the running example. *COH* = cohesion; *LS* = leadership style; *SAT* = job satisfaction; *WL* = workload.

intercept $\mathbf{u}_{1j} = (u_{SAT,j}, u_{LS,j}, u_{WL,j})$, representing the components of these variables that vary between groups, and a residual term $\mathbf{r}_{1ij} = (r_{SAT,ij}, r_{LS,ij}, r_{WL,ij})$, representing the differences within groups. For cohesion, which was assessed directly at Level 2, the model includes a residual term $\mathbf{u}_{2j} = (u_{COH,j})$. The critical point in this model is that it assumes that the random effects and residuals at Level 2 (i.e., global and shared variables) may be correlated ($\boldsymbol{\Psi}$) and that the residuals at Level 1 may be correlated as well ($\boldsymbol{\Sigma}$). This illustrates that the joint model indeed "matches" the multilevel structure because it allows the user to differentiate between (a) the within- and between-group components that can be present in variables at Level 1 and (b) the relations between variables within and between groups. The joint model (or variants thereof) is implemented in the packages pan (Schafer & Zhao, 2016) and jomo (Quartagno & Carpenter, 2016) for the statistical software R as well as in the standalone software packages SAS (Mistler, 2013), M*plus* (Asparouhov & Muthén, 2010), and REALCOM (Carpenter, Goldstein, & Kenward, 2011).

Fully conditional specification. As an alternative to the joint model, the joint distribution of the variables with missing data can be approximated by imputing one variable at a time using a sequence of univariate models. To address multivariate patterns of missing data, the procedure iterates back and forth between variables with missing data, conditioning on the other variables in the data set (or a subset of them). This approach is referred to as the fully conditional specification of MI (FCS; van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006). Specifically, for a set of variables at Levels 1 and 2, a sequence of conditional imputation models can be specified as follows:

$$y_{1ijp} = \mathbf{y}_{1ij(-p)}\boldsymbol{\gamma}_p + u_{jp} + r_{ijp} \qquad \text{(Level 1)}$$

$$y_{2jq} = \mathbf{y}_{j(-q)}\boldsymbol{\gamma}_q + u_{jq}, \qquad \text{(Level 2)} \quad (16.5)$$

where y_{1ijp} is the p-th variable with missing data at Level 1, and $\mathbf{y}_{ij-(p)}$ is a set of predictors for that variable that may include any variable other than y_{1ijp}. Similarly, y_{2jq} is the q-th variable with missing data at Level 2 (i.e., a global variable), and $\mathbf{y}_{j-(q)}$ is a set of predictor variables that may include any other variable at Level 2 (i.e., global variables) as well as the between-group components of any variable at Level 1. The random intercepts u_{jp} as well as the residuals r_{ijp} and u_{jq} in each model are each assumed to follow independent normal distributions (see also van Buuren, 2011). To address multiple variables with missing data, the FCS algorithm arranges them in a sequence and visits one variable at a time, generating imputations from the imputation model assigned to each variable. Once a variable has been completed in this manner, it can be used as a predictor in any of the other imputation models. After each variable has been visited, the sequence is repeated, and new imputations are generated until the algorithm converges, yielding the first of multiple imputations.

The sequential nature of the FCS algorithm requires some rethinking. In contrast to the joint model, the FCS algorithm allows different predictors to be selected for each target variable, and conversely, all target variables can act as predictors in any other target's imputation model. Moreover, in order to preserve the relationships between variables, it is in fact *required* that the imputation model for each target variable is conditioned on the other variables. To incorporate relationships between variables at Level 2, the group means of the variables at Level 1 must be calculated and included as predictors. In addition, the group means must be updated once new imputations for the underlying variables have been obtained; this process of updating the group means is known as *passive* imputation (e.g., Royston, 2005).

To illustrate multilevel FCS, consider our running example and the illustration in Figure 16.5. Missing data in job satisfaction, leadership style, and workload can be imputed by applying separate multilevel models, where the model for workload should be appropriate for binary categorical data (e.g., a logistic multilevel model). Cohesion can be imputed by using a regression model at Level 2. To preserve the relationships between the variables within and between groups, all variables are included as predictor variables in the other variables' imputation models, and the group means are updated and included by using passive imputation. The FCS and similar approaches for multilevel data are implemented in the package mice (van Buuren & Groothuis-Oudshoorn, 2011) for the statistical software R as well as in the standalone software packages M*plus* (Asparouhov & Muthén, 2010) and Blimp (Keller & Enders, 2018).

Incomplete categorical variables. There are several options for treating missing values in categorical and ordinal variables. The first option is to treat categorical variables as continuous for the purpose of MI and to round the resulting values to comply with the original categories in that variable. For example, imputations for ordinal data may be rounded using 0.5, 1.5, and so forth as thresholds; for binary data, adaptive rounding can be applied, which uses the mean of the imputed values to adjust the threshold accordingly (see Carpenter & Kenward, 2013). Adaptive rounding has been shown to perform well for binary missing data (Bernaards, Belin, & Schafer, 2007), but also MI without rounding appears to work well for binary and (some) ordinal variables (Schafer, 1997; W. Wu, Jia, & Enders, 2015). Finally, it is possible to impute categorical and ordinal variables using a latent variable approach. In this approach, imputations are generated for a set of underlying latent variables that represent the relative probability of being assigned to a category. Based on the latent scores, the assignment to a category can then be simulated by using an appropriate link function (e.g., a probit link for latent normal variables; see Carpenter & Kenward, 2013). For a variable with C categories, this approach introduces $C - 1$ latent variables that represent the possible contrasts between categories (Carpenter & Kenward, 2013; see also Goldstein et al., 2009). For binary variables, this is equivalent to generating imputations from a generalized linear mixed-effects model (e.g., a logistic or probit model). These procedures, too, appear to work well for both

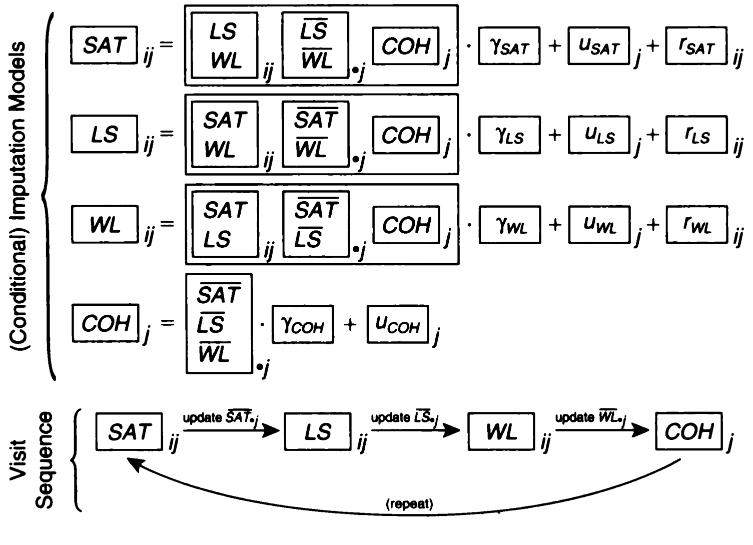
(Conditional) Imputation Models
SAT ij = LS WL ij, LS WL •j, COH j · γSAT + uSAT j + rSAT ij
LS ij = SAT WL ij, SAT WL •j, COH j · γLS + uLS j + rLS ij
WL ij = SAT LS ij, SAT LS •j, COH j · γWL + uWL j + rWL ij
COH j = SAT LS WL •j · γCOH + uCOH j
Visit Sequence
SAT ij
update SAT•j
LS ij
update LS•j
WL ij
update WL•j
COH j
(repeat)

FIGURE 16.5. Schematic representation of the sampling steps in the fully conditional specification of multilevel multiple imputation in the running example. *COH* = cohesion; *LS* = leadership style; *SAT* = job satisfaction; *WL* = workload.

binary and polytomous data (Demirtas, 2009; W. Wu et al., 2015; see also Enders et al., 2016).

Analyzing multiply imputed data. The idea underlying MI is to generate plausible replacements for each missing value, thus transforming a data set with "missing data" into a data set with "complete data." This process is repeated M times (hence, the qualifier "multiple"), yielding M completed versions of the original data (see Figure 16.2). Once the set of M data sets has been obtained, the model of interest must be fit separately to each data set, yielding M estimates of some parameter of interest, say $\hat{Q}_m$ (e.g., regression coefficients; $m = 1, \ldots, M$), and M estimates of the sampling variance of that estimate, $\hat{V}_m$ (e.g., squared standard errors). According to Rubin (1987), the combined point estimate is the average of the individual estimates

$$\hat{Q}_{MI} = \frac{1}{M}\sum_{m=1}^{M}\hat{Q}_m. \tag{16.6}$$

The combined estimate of the sampling variance of the estimator incorporates two different sources of uncertainty:

$$\hat{V}_{MI} = \hat{W} + \left(1 + \frac{1}{M}\right)\hat{B}, \tag{16.7}$$

where $\hat{W}$ denotes the sampling variance *within* imputations, that is, the average of the individual variance estimates

$$\hat{W} = \frac{1}{M}\sum_{m=1}^{M}\hat{V}_m, \tag{16.8}$$

and $\hat{B}$ denotes the sampling variance *between* imputations, that is, the variance of the point estimates across data sets:

$$\hat{B} = \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{Q}_m - \hat{Q}_{MI}\right)^2. \tag{16.9}$$

Using the combined point and variance estimates, standard hypothesis tests can be carried out on the basis of a Student's t distribution with v degrees of freedom. Rubin (1987) recommended calculating the degrees of freedom as follows:

$$v = (M-1)\left[1 + \frac{1}{\mathrm{RIV}}\right]^2, \tag{16.10}$$

where the expression

$$\mathrm{RIV} = \frac{\hat{W}}{(1 + 1/M)\hat{B}}, \tag{16.11}$$

denotes the relative increase in the sampling variance of the estimator that is due to missing data (see also Barnard & Rubin, 1999). In addition, several alternative formulas have been proposed for more complex hypotheses that may involve several parameters simultaneously, for example, when testing the overall effect of categorical explanatory variables or when testing for random slopes using a likelihood-ratio test (see Appendix 16.1; see also Reiter & Raghunathan, 2007).

The general idea behind Rubin's rules is to approximate the sampling distribution of $\hat{Q}$ that would be obtained with infinite M but based on only a small number of imputations. Naturally, the larger the number that is chosen for M, the better the approximation becomes, which raises the question of "How many are needed?" Traditionally, $M = 5$ imputations have been recommended (Rubin, 1987), but more can be necessary when the amount of missing data increases or the model of interest becomes more complex (Bodner, 2008; Graham, Olchowski, & Gilreath, 2007). This is especially important because most software packages for multilevel MI generate $M = 5$ imputations by default. In our experience, $M = 20$ imputations are usually sufficient for estimating and testing the parameters in most applications of multilevel models. However, when large portions of the data are missing (say more than 50%) or complex hypotheses that involve multiple parameters are being tested, we recommend generating 50 to 100 imputed data sets (see also Bodner, 2008; Raghunathan, 2015).

Maximum Likelihood

The general principle behind ML estimation is that the values of the parameters in a statistical model can be chosen in such a way that the likelihood of the data becomes maximal. When the data contain missing values, it is often possible to estimate the model directly using only the observed data. This procedure is often referred to as *direct* or *full information* ML. Using ML, the likelihood is evaluated on a case-by-case basis; that is, cases with incomplete records contribute to the likelihood only to the extent to which they have data (Little & Rubin, 2002). The ML estimates of the parameters in a model of interest are consistent when the data are MAR or MCAR; that is, missing data occur in an unsystematic fashion when the variables in the model are taken into account (Little & Rubin, 2002).

The main principle by which ML "deals" with missing data is that it imposes distributional assumptions on incomplete variables. For this reason, common multilevel software packages often handle missing values only in the dependent variable of the model (e.g., HLM, SAS), where such assumptions are already in place, but cases with missing values in explanatory variables are discarded because no distributional assumptions have been made for them. To circumvent this restriction, it has been suggested that researchers adopt the framework of *structural equation modeling* (SEM), which allows the user to introduce distributional assumptions for all variables by defining them as endogenous (i.e., dependent) variables in a single analysis model (e.g., Allison, 2012; Enders, 2010). For example, in the statistical software M*plus*, this is achieved by including the variances and covariances of the explanatory variables in the modeling statement. Using this strategy, it is often possible to prevent the software from discarding these cases and to apply the ML principle to both the dependent and explanatory variables in a model of interest. Furthermore, this perspective offers the possibility of including auxiliary variables that may improve the plausibility of the MAR assumption and the accuracy of estimates under ML (Enders, 2008; Graham, 2003). Software that supports ML for multilevel models from the perspective of SEM includes the standalone software packages M*plus* (Muthén & Muthén, 2012), Latent GOLD (Vermunt & Magidson, 2016), gllamm (Rabe-Hesketh, Skrondal, & Pickles, 2004), and xxM (Mehta, 2013).

As an alternative to direct ML, estimates of the parameters in a multilevel model can be obtained from a two-stage procedure by first estimating a covariance matrix within and between groups on the basis of the observed data; in the second stage, the parameters of interest are derived from the variances and covariances estimated in the first stage (Yuan & Bentler, 2000). Conceptually, two-stage ML is similar to the perspective taken in SEM. We will not consider this approach further, but using two-stage ML can offer advantages when working with nonnormal variables and because auxiliary variables are easily incorporated into the estimation procedure (Savalei & Bentler, 2009; Yuan, Yang-Wallentin, & Bentler, 2012).

Comparison of Maximum Likelihood and Multiple Imputation

From a theoretical point of view, ML and MI are not vastly different, and the two can be expected to yield similar results when they operate under similar assumptions (Schafer & Graham, 2002). However, from a practical point of view, the differences can be substantial. Fitting models using ML is often easy, provided that a software package that supports the estimation of the model of interest can be found. Furthermore, because ML does not separate the treatment of missing data from the analysis, the missing data model is always consistent with the analysis model; that is, the two models are always based on the same set of assumptions (Allison, 2012). However, integrating the treatment of missing data and the estimation of the analysis model into a single step also has disadvantages. First, the distributional assumptions needed for the treatment of missing data now also enter the analysis model even though they might not have originally been part of it. Second, auxiliary variables must be incorporated directly into the model of interest, thus making the analysis model more complex (Graham, 2003). In applications with only a few well-behaved variables, this is usually not a problem; but in practice, it can become problematic, for example, when the inclusion of auxiliary variables leads to a mix of continuous and categorical variables at both Levels 1 and 2. Such models are difficult for the user to specify, and a given software package might not even fully support it, forcing the user to alter the model or make decisions he or she would not have made otherwise.

Conducting MI, on the other hand, is more complicated at first glance. First, an imputation model that is consistent with the model of interest must be chosen. Then, the user must specify the number of imputations and the number of iterations for which the sampling procedure should run. Finally, he or she must ensure that the algorithm has converged before any analyses can be carried out (see also Allison, 2012). Once the imputations have been generated, the user must fit the analysis model to each of the imputed data sets and combine their results into a final set of parameter estimates and inferences. Especially for inexperienced users, performing MI can be a daunting task. On the other hand, modern procedures for multilevel MI are powerful and very flexible in accommodating a variety of models. In addition, many software packages for multilevel MI automatize at least some of these steps. Finally, the separation between the treatment of missing data and the analysis phase makes it straightforward to handle a variety of variables and to include auxiliary variables without altering the model of interest.

SIMULATION

Next, we report the results from a computer simulation study. This study was intended to illustrate the general performance of ML and MI in a controlled setting. We conducted this study with two models of interest in mind. The first model of interest (Model 1) was the model from our running example:

$$\begin{aligned} SAT_{ij} = \gamma_{00} + \gamma_{10}\left(LS_{ij} - \overline{LS}_{\cdot j}\right) + \gamma_{01}\overline{LS}_{\cdot j} \\ + \gamma_{20}WL_{ij} + \tilde{a}_{02}COH_{j} + u_{0j} + r_{ij}. \end{aligned} \tag{16.3, revisited}$$

This represents the standard formulation of multilevel models in which the observed group means represent the shared perception of leadership style among members of the same group.

The second model of interest (Model 2) is also known as the "multilevel latent covariate model" (Lüdtke et al., 2008) and differs from the first model in that it uses the true, unobserved group means or between-group components to represent the shared perception of individuals in each group. The model reads

$$\begin{aligned} SAT_{ij} = \gamma_{00} + \gamma_{10}LS_{W.ij} + \gamma_{01}LS_{B,j} + \gamma_{20}WL_{ij} \\ + \gamma_{02}COH_{j} + u_{0j} + r_{ij}, \end{aligned} \tag{16.12}$$

where $LS_{W,ij}$ and $LS_{B,j}$ denote the within- and between-group components of leadership style (Asparouhov & Muthén, 2006; Lüdtke et al., 2008). Formulating the model in terms of the true within- and between-group components can be beneficial because it corrects for the fact that the group mean is calculated from a finite number of observations and thus provides only an unreliable measure of the true between-group component (see Croon & van Veldhoven, 2007; Raudenbush & Bryk, 2002). In the organizational literature, the reliability of the group mean is also known as the ICC(2), and it expresses the extent to which differences between the observed group means reflect true differences between groups (Bliese, 2000; see also LeBreton & Senter, 2008). It is a matter of debate in the multilevel literature which formulation of the model of interest is more appropriate. For example, it can be argued that the formulation in Model 2 is appropriate if the shared perception among individuals is of primary interest (e.g., ratings of team climate, leadership effectiveness), whereas Model 1 may be appropriate if the variation within groups is itself of interest or if the observed group mean is simply regarded as a summary measure (e.g., gender ratio, socioeconomic status; for further discussion, see Lüdtke et al., 2008). However, the main motivation for including these two approaches to modeling between-group effects in the present chapter was that their distinction is important for the treatment of missing data under ML (see below).

In the simulation study, the samples were generated from either Model 1 (the "standard" model) or Model 2 (the "latent covariate" model) in order to allow for a comparison between conditions in which one of the two models is the "true" model. The parameters of the simulation were loosely based on the data from Klein et al. (2000). The samples consisted of $G = 50$ groups of size $n = 10$. All variables were standardized across groups with mean zero and unit total variance. For the ratings on leadership style and job satisfaction, we assumed ICCs of .10 and .20, respectively. In addition, we assumed that negative leadership style was correlated with cohesion at the group level ($r = -.15$). For the two workload categories (high vs. low), we generated a standard normal variable with an ICC of .20, and we used 0.38 as a breaking point to dichotomize that variable, resulting in 35% and 65% of individuals with high and low workloads, respectively. For simplicity, we assumed that workload was uncorrelated with the other explanatory variables. Finally, we assumed the following fixed effects in the data-generating model: $\gamma_{00} = 0$ (intercept), $\gamma_{10} = -.20$ and $\gamma_{01} = -.70$ (leadership style), $\gamma_{20} = -.30$ (workload), and $\gamma_{02} = .10$ (cohesion). The variance components τ_0^2 and σ^2 then followed. We induced missing values in cohesion completely at random (5%) and in leadership style (15%) and workload (10%) on the basis of job satisfaction (lower job satisfaction corresponded with a greater chance of missing data). Finally, we induced missing values in job satisfaction completely at random (10%).

Using this procedure, we generated 5,000 data sets from both Models 1 and 2. In each data set, we carried out MI using both joint modeling (using jomo; Quartagno & Carpenter, 2016) and FCS (using mice; van Buuren & Groothuis-Oudshoorn, 2011) in the statistical software R. Afterwards, we fitted the respective model of interest using M*plus* 7 (Muthén & Muthén, 2012). We also used M*plus* to estimate the model with ML, and we addressed missing data in explanatory variables by specifying distributional assumptions for these variables. In the context of Model 2, applying ML is relatively easy because M*plus* already imposes the necessary distributional assumptions when decomposing leadership style into its within- and between-group components. The distributional assumptions for the remaining variables can be added by defining them as endogenous variables at Level 1 or Level 2, respectively.[2] On the other hand, in the context of Model 1, missing data in explanatory variables posc a greater challenge when estimating the model using ML. We consider two strategies for this case, neither of which is completely satisfying. In the first strategy (ML1), distributional assumptions are specified as before by defining explanatory variables as endogenous variables at Levels 1 and 2, respectively. However, this strategy unintentionally adopts the within- and between-group decomposition for leadership style (as in Model 2), thus correcting between-group effects that did not require correction. As a second option (ML2), the group means of leadership style can be calculated beforehand from the observed data, and distributional assumptions can be imposed only on the within-group deviations of leadership style. In this specification, the group means are consistent with the analysis model, but the between-group effects of leadership style may be biased if values are missing in a systematic manner (similar to LD).

In Table 16.2, we included the mean estimates of the three procedures for the two models of interest as well as the coverage of the 95% confidence interval. Ideally, the mean estimates should be close to the true values in the data-generating model, and the coverage rates should be close to 95%. In the context of Model 2, both MI and ML yielded parameter estimates that were very close to the true values, and coverage rates were close to the nominal value of 95%. However, the between-group effect of leadership style (γ_{01}) was slightly too large under ML, which may be attributed to the small sample size at Level 2 (Lüdtke et al., 2008). In the context of Model 1, the parameter estimates obtained from MI were again close to the true values, but the between-group effect of leadership style (γ_{01}) was slightly underestimated. Under ML, specifying leadership style as an endogenous variable (ML1), and thus adopting the within- and between-group decomposition, led to severe bias in the between-group regression coefficients. By contrast, when the group means were calculated beforehand

[2] Using ML, it was also not straightforward to accommodate both (a) the multilevel structure of the variables and (b) the fact that workload is categorical. Therefore, we treated workload as a continuous variable. Although this may be acceptable for a dichotomous variable with similar frequencies in both categories, it will lead to problems when explanatory variables have multiple categories or some categories occur much more frequently than others.

TABLE 16.2

Mean Estimates (and Coverage Rates for the 95% Confidence Interval) for the Two Models of Interest for Multiple Imputation and Maximum Likelihood

	Model 1					Model 2			
	True	JM	FCS	ML1	ML2	True	JM	FCS	ML1
γ_{00}	0.000	0.003	0.002	0.004	0.011	0.000	0.001	0.000	0.001
		(95.0)	(95.0)	(96.1)	(94.1)		(94.8)	(94.3)	(95.1)
γ_{10}	−0.200	−0.202	−0.200	−0.203	−0.200	−0.200	−0.201	−0.200	−0.200
		(94.7)	(94.7)	(94.9)	(94.7)		(93.8)	(94.0)	(94.3)
γ_{01}	−0.700	−0.648	−0.660	−1.215	−0.633	−0.700	−0.708	−0.714	−0.803
		(94.6)	(94.5)	(91.8)	(90.5)		(95.4)	(95.1)	(96.1)
γ_{20}	−0.300	−0.301	−0.300	−0.302	−0.303	−0.300	−0.298	−0.297	−0.299
		(94.8)	(95.0)	(94.9)	(94.8)		(94.9)	(94.9)	(94.9)
γ_{02}	0.100	0.102	0.102	0.067	0.105	0.100	0.098	0.099	0.095
		(94.3)	(93.8)	(95.5)	(92.7)		(95.0)	(94.4)	(95.4)
τ_0^2	0.083	0.085	0.082	0.035	0.081	0.088	0.079	0.078	0.069
		(95.4)	(94.0)	(91.9)	(91.5)		(95.0)	(93.9)	(92.3)
σ^2	0.751	0.747	0.747	0.746	0.749	0.790	0.786	0.786	0.786
		(94.1)	(94.1)	(94.1)	(94.3)		(94.6)	(94.5)	(94.8)

Note. FCS = fully conditional specification of multiple imputation; γ_{00} = intercept; γ_{10} = within-group effect of leadership style; γ_{01} = between-group effect of leadership style; γ_{20} = effect of workload; γ_{02} = effect of cohesion; JM = joint modeling of multiple imputation; ML1 = maximum likelihood with true within- and between-group components for leadership style; ML2 = maximum likelihood with group means for leadership style calculated from the observed data; σ^2 = residual variance; τ_0^2 = intercept variance.

from the observed data (ML2), thus treating only the within-group deviations as endogenous, the group-level effect of leadership style (γ_{01}) was only slightly underestimated. The coverage rates were relatively close to the nominal value of 95% for most parameters but tended to be slightly smaller under ML, especially when the group means were calculated from the observed data (ML2).

In conclusion, both ML and MI provided accurate results when their assumptions were met and when these assumptions were consistent with the model of interest. These requirements were more easily fulfilled in the context of Model 2, in which case both MI and ML yielded reasonable parameter estimates. However, in the context of Model 1, the results were more diverse. Under ML, following the usual advice to treat explanatory variables as endogenous can lead to an unwanted "shift" in the analysis model; in the present case, this resulted in parameter estimates that were severely distorted. When the group means were calculated beforehand, we observed only little bias. However, this approach slightly overestimated the precision of the parameter estimates because it ignored the fact that group means were calculated from incomplete records. Under MI, estimates were accurate, and the confidence intervals showed good coverage properties, providing the most reasonable approximation to the true parameters overall.

EXAMPLE APPLICATION

In this section, we apply the missing data methods to our running example. The running example is based on the data from Klein et al. (2000) and essentially mimics the conditions in our simulation study except that the example data set contains *un*standardized variables instead. Missing values were induced in the data set in the same way as in the simulation study. As a result, 21.9% of the employees had missing values on at least one variable; these were distributed across job satisfaction (9.2%), leadership style (12.3%), workload (11.5%), and cohesion (4.0%).

The data set is included in the R package mitml (Grund, Robitzsch, & Lüdtke, 2016). The model of interest was the "standard" multilevel model in Equation 16.3 (Model 1). We applied MI using the joint model implemented in the jomo package in R, and we estimated the model of interest using the lme4 package (Bates, Maechler, Bolker, & Walker, 2016). To assist with the analyses, we used the mitml package, which provides a wrapper function for the jomo package as well as tools for analyzing multiply imputed data sets (see also Grund et al., 2016b). For ML estimation, we used M*plus*, where we calculated the group means of leadership style from the observed records (as in ML1) and adopted the within- and between-group decomposition for the remaining variables (as in ML2). The computer code and the M*plus* syntax file are provided in Appendix 16.2.

To set up the imputation model using jomo and mitml, two formulas that denoted the imputation model for variables at Levels 1 and 2, respectively, had to be specified (see Equation 16.4 and Figure 16.4). In accordance with the "empty" specification of the model, all variables were treated as target variables, and no predictor variables were specified except for a "one" for the intercept. We generated $M = 100$ imputations in this manner. The number of iterations for the algorithm was chosen in such a way that convergence could be established by inspecting convergence criteria (e.g., Gelman & Rubin, 1992) and diagnostic plots for the parameters of the imputation model (Grund et al., 2016b; see also Schafer & Olsen, 1998). After running MI, the model of interest was fitted to each of the imputed data sets using lme4, and the parameter estimates were pooled by employing Rubin's rules in order to obtain a final set of parameter estimates and inferences. The results obtained from ML and MI are presented in Table 16.3. The two analyses suggested that negative leadership style had a relatively strong impact on employees' job satisfaction when employees' workload and the work group's cohesion were taken into account. Under MI, for any one-unit change in the leadership style ratings within groups (Level 1), the expected change in job satisfaction was −.532 ($p < .001$). Between groups, a one-unit change in the shared perception of leadership style ratings (Level 2) was associated with an expected change in job satisfaction of −1.566 ($p < .001$).

TABLE 16.3

Estimates for the Parameters in the Model of Interest Obtained From Maximum Likelihood and Multiple Imputation in the Running Example

	M*plus* (ML2)		jomo (MI)			
Parameter	**Est.**	**SE**	**Est.**	**SE**	**RIV**	**FMI**
Intercept (γ_{00})	0.291^{***}	0.136	$0.257^{\dagger**}$	0.140	0.167	0.143
Level 1						
Leadership style (γ_{10})	-0.526^{***}	0.091	-0.532^{***}	0.092	0.341	0.255
Workload (γ_{20})	-0.863^{***}	0.197	-0.842^{***}	0.195	0.259	0.206
Level 2						
Leadership style (γ_{01})	-1.491^{***}	0.319	-1.566^{***}	0.349	0.237	0.192
Cohesion (γ_{02})	0.237^{***}	0.088	0.243^{***}	0.091	0.075	0.070
Level 2 residual variance (τ_0^2)	0.268^{***}	0.128	0.286^{***}			
Level 1 residual variance (σ^2)	4.940^{***}	0.283	4.962^{***}			

Note. ML2 = maximum likelihood with group means for leadership style calculated from the observed data; MI = multiple imputation; Est. = estimate; SE = standard error; RIV = relative increase in variance; FMI = fraction of missing information; γ_{00} = intercept; γ_{10} = within-group effect of leadership style; γ_{20} = effect of workload; γ_{01} = between-group effect of leadership style; γ_{02} = effect of cohesion; τ_0^2 = intercept variance; s_2 = residual variance.

$^{\dagger}p < .10$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$ (two-tailed).

Furthermore, there was a negative effect of high (vs. low) workload (−0.842, $p <. 001$) on job satisfaction and a positive effect of cohesion (0.243, $p = 0.007$). The results obtained from ML were virtually identical. Perhaps the largest difference between the two procedures was the standard error for the between-group effect of leadership style, which might reflect the slightly too narrow confidence intervals under ML observed in the simulation study.

In addition, we also investigated whether the within-group effect of leadership style *varied* across groups, that is, whether there was significant variance in the slope of leadership style. To this end, we fitted an alternative model that contained a random slope for within-group effect of leadership style. The alternative model was compared with the model of interest using the D_3 statistic (Meng & Rubin, 1992), which can be interpreted as a pooled LRT for multiply imputed data sets (Appendix 16.1). The D_3 statistic suggested that there was not enough evidence to conclude that the effect of leadership style truly varied across groups, $F(2,3707.9) = 2.621$ ($p = .071$). Therefore, the alternative model was rejected in favor of the model of interest.[3] Furthermore, we were interested in whether the effect of leadership style was larger between than within groups. For this purpose, we used the D_1 statistic, which allowed us to test the difference between the two coefficients against zero by specifying it as a linear constraint (Appendix 16.1; see also Kreft, de Leeuw, & Aiken, 1995). The D_1 statistic suggested that the two parameters were significantly different from one another, $F(1,2471.3) = 8.253$ ($p = .004$), that is, the between-group effect (−1.57) was significantly larger than the within-group effect (−0.53).

DISCUSSION

In this chapter, we provided an introduction to multilevel modeling with missing data. In particular, we looked at two principled methods for handling missing data: MI and estimation by ML. The general ideas behind ML and MI are not vastly different, and both may be regarded as state-of-the-art procedures for handling missing data (Schafer & Graham, 2002). The differences between the two methods are most often of a practical nature. Although the two procedures tend to give the same answers if they are based on similar assumptions, carrying out a given task is often easier with one procedure than the other. For example, ML is very easy to incorporate into one's regular workflow because the missing data treatment is performed during the estimation of the model of interest (see also Allison, 2012). On the other hand, addressing missing values and including auxiliary variables may prove to be challenging depending on where the missing data occur and how complex the model becomes once all factors are taken into account, for example, if categorical variables contain missing data or between-group effects are represented by observed group means. By contrast, MI allows for the very flexible modeling of different types of variables and including auxiliary variables is straightforward. On the other hand, performing MI and analyzing multiple data sets can be challenging, especially for less experienced users or when nonstandard analyses and hypothesis tests are required. That being said, although we clearly see ML as the easier-to-use alternative (see Allison, 2012; Enders, 2010), we tend to favor MI for its flexibility and because it separates the imputation from the analysis phase (see Carpenter & Kenward, 2013; Schafer & Graham, 2002; see also Grund, Lüdtke, & Robitzsch, 2018).

As in every introduction to these or similar procedures, it is not possible to give all possible research scenarios the attention they deserve. In this chapter, we restricted our discussion to cross-sectional multilevel models with a single level of clustering, that is, individuals nested within some higher-level collective. In principle, the procedures discussed here generalize naturally to models with additional levels of clustering, for example, three-level models (Goldstein, 2011; Keller, 2015; Yucel, 2008), models with cross-classified random effects (Goldstein, 2011; Hill & Goldstein, 1998), or models with multiple memberships (Goldstein,

[3] Note that, because the imputation model did not include random slopes, it did not "match" the alternative model. For this reason, the hypothesis test' was not completely trustworthy and was included here only for the purpose of illustration.

2011; Yucel, Ding, Uludag, & Tomaskovic-Devey, 2008). However, these procedures are not widely available in standard software, and more research is needed to evaluate their performance in realistic research scenarios.

Another topic that we did not discuss explicitly is the treatment of missing data in longitudinal research designs (e.g., repeated measurements, diary studies, experience sampling, ecological momentary assessment). This topic is particularly interesting, however, because multilevel models are frequently used to analyze longitudinal data. Fortunately, many of the ideas presented here can also be applied to longitudinal data (see also Black, Harel, & Matthews, 2013; Newman, 2003). For example, assume that a researcher is interested in estimating a growth curve model with missing data in the dependent variable that should be treated using MI. It is then useful to distinguish studies in which the longitudinal design is balanced or unbalanced with respect to time, that is, whether all participants were measured at the same or a different set of time points (see W. Wu, West, & Taylor, 2009). If all participants were measured on the same set of time points, then the longitudinal data structure can be expressed in a wide data format, and single-level MI can be used to treat the missing values in the dependent variable (for a two-stage ML procedure, see Yuan et al., 2012). However, if participants were measured at potentially different or unbalanced time points, then procedures based on mixed-effects models for multilevel MI may be more appropriate (see Equation 16.4). However, even though the model by Schafer and Yucel (2002) was developed explicitly with applications to longitudinal data in mind, the model lacks the flexibility to incorporate some covariance structures at Level 1 that are commonly used in longitudinal analysis models (see Pinheiro & Bates, 2000). Similar problems may be observed when ML is used to estimate growth curve models because it is difficult to establish a homogeneous covariance structure for this type of data (W. Wu et al., 2009).

Even though there has been a substantial amount of interest in missing data methods for multilevel data in recent years, some questions still provide challenges for the future. One such example is the treatment of missing data in multilevel models with random slopes or in models with nonlinear and interaction effects. For example, it has been shown that current implementations of MI are not perfectly suited for handling missing data in explanatory variables in multilevel models with random slopes (e.g., Enders et al., 2016; Gottfredson, Sterba, & Jackson, 2017; Grund, Lüdtke, & Robitzsch, 2016a; see also von Hippel, 2009). Similar problems may occur under ML but have yet to be discussed more thoroughly in the applied missing data literature (however, see Enders et al., 2014). In order to make sure that imputations are consistent with the model of interest, it has been argued that the substantive analysis model should be taken into account during MI (Bartlett, Seaman, White, & Carpenter, 2015; Carpenter & Kenward, 2013). Several authors have proposed procedures that incorporate these ideas using rejection sampling or a Metropolis–Hastings algorithm for multilevel MI, but these procedures are not yet available in standard software (Erler et al., 2016; Goldstein, Carpenter, & Browne, 2014; L. Wu, 2010). Similar procedures have been proposed in the context of ML, where the likelihood function in a multilevel model can be factored into separate components that refer to the model of interest and additional models for explanatory variables with missing data (Ibrahim, Chen, & Lipsitz, 2001; Stubbendick & Ibrahim, 2003).

To sum up, missing data are an ever-present problem in research practice. We believe that ML and MI provide powerful tools for the treatment of missing data in multilevel research. The two procedures both come with their own strengths and weaknesses, and one may be preferred over the other for a specific missing data problem. At the end of the day, however, they are more similar than they are different, and both offer a substantial improvement over approaches such as LD in terms of generality, theoretical foundation, accuracy of parameter estimates, and statistical power. In the present chapter, we provided an introduction to these methods, and we offered guidance on how to apply them in multilevel research. The treatment of missing data is not without its challenges, and there remain many open (and interesting) questions for the future.

However, we believe that these methods are a valuable addition to the researcher's toolbox, which, if applied correctly, can improve the quality of the conclusions we draw from our data and that of our research altogether. We hope that this chapter will promote the adoption of MI and ML and will encourage researchers to use these procedures in their own research projects.

APPENDIX 16.1: MULTIPARAMETER HYPOTHESIS TESTS IN MULTIPLE IMPUTATION

In research practice, statistical hypotheses often involve multiple parameters simultaneously (e.g., linear constraints, comparisons of nested models). In analyses with complete data, multiparameter hypothesis tests are often performed using the Wald test or likelihood-ratio test (LRT). To pool a series of Wald tests on the basis of a series of parameter vectors, $\hat{Q}_m$, and covariance matrices, $\hat{V}_m$, Li, Raghunathan, and Rubin (1991) proposed that researchers should use the test statistic

$$D_1 = \frac{\left(\hat{Q}_{MI} - Q_0\right)^T \hat{W}^{-1}\left(\hat{Q}_{MI} - Q_0\right)}{K(1 + \mathrm{ARIV}_1)}, \quad (16A.1)$$

where $\hat{Q}_{MI}$ and $\hat{W}$ are the average estimates of the parameter vector and its covariance matrix (see Equations 16.6 and 16.8), Q_0 contains the hypothesized values of the parameters under the null hypothesis, and ARIV_1 is an estimate of the average relative increase in variance (ARIV) due to nonresponse across parameters (see Enders, 2010). The D_1 statistic can be used in a similar manner as Rubin's rules (1987), that is, it can be used to test a set of parameters (or a linear transformation thereof) that have an approximately normal sampling distribution (e.g., regression coefficients).

It is sometimes difficult to calculate D_1, for example, because estimates of the covariance matrix are unavailable. As an alternative, Li, Meng, Raghunathan, and Rubin (1991) proposed that a set of Wald-like test statistics, D_m, be pooled as follows:

$$D_2 = \frac{\bar{D}K^{-1} + (M+1)(M-1)^{-1}\mathrm{ARIV}_2}{1 + \mathrm{ARIV}_2}, \quad (16A.2)$$

where $\bar{D}$ is the average of the D_m, and ARIV_2 is an alternative estimate of the ARIV. The D_2 statistic can be used for any quantity that follows a χ^2-distribution, for example, a Wald test of a set of regression coefficients (or a linear transformation thereof) or an LRT comparing two nested models (see also Snijders & Bosker, 2012).

As a third option, Meng and Rubin (1992) proposed a test statistic for pooling a series of LRTs as follows:

$$D_3 = \frac{\tilde{L}}{K(1 + \mathrm{ARIV}_3)}, \quad (16A.3)$$

where the ARIV_3 is another estimate of the average relative increase in variance, which includes (a) the average LRT statistic evaluated at the *actual* parameter estimates and (b) the average LRT statistic evaluated at the *average* parameter estimates for the two models ($\tilde{L}$). This test statistic can be used in the same manner as the LRT, for example, for comparing two nested statistical models (see above).

In general, D_1 and D_3 tend to be the more reliable procedures and should be used when possible. However, because software implementations of D_1 and D_3 are sometimes not available, D_2 may be an interesting alternative given its ease of application. Even though D_2 was optimized to work with a small number of imputations ($M = 3$), results from D_2 tend to be much more robust when more imputations (say, $M \geq 20$) are used (Grund, Lüdtke, & Robitzsch, 2016c; Licht, 2010). Care should be taken when large portions of the data are missing (say, more than 50%) because D_2 and (to a lesser extent) D_3 tend to be less robust in these cases.

APPENDIX 16.2: COMPUTER CODE FOR THE EXAMPLE APPLICATION

Printed below is the computer code used for multilevel MI in the data analysis example.

```
# *** Description of the 'leadership' data set:
#
# GRPID: indicator for work groups
# JOBSAT: job satisfaction (Level 1)
# NEGLEAD: ratings on negative leadership
   style (Level 1)
# WLOAD: workload (Level 1, "low" vs. "high")
# COHES: group cohesion (Level 2)

# Multiple imputation is performed with an "empty"
   joint model using jomo. The
# model of interest is fit using lme4, and the mitml
   package is used for pooling
# tests and parameters.

library(lme4)
library(mitml)

# set up random number generator
set.seed(1234)

# load data
data(leadership)

# *** Imputation phase:
#
# set up "empty" model
fml <- list(NEGLEAD + JOBSAT + WLOAD ~ 1
   + (1|GRPID),   # Level 1 model
   COHES ~ 1)   # Level 2 model

# impute
imp <- jomoImpute(leadership, formula=fml,
   n.burn=5000, n.iter=500, m=100)

# assess convergence
summary(imp)   # convergence criteria ("Rhat")
plot(imp)   # diagnostic plots

# create list of completed data sets
implist <- mitmlComplete(imp, print="all")

# *** Analysis phase:
#

# apply group-mean centering
implist <- within(implist,{
   G.NEGLEAD <- clusterMeans(NEGLEAD,GRPID)
   I.NEGLEAD <- NEGLEAD - G.NEGLEAD
})

# fit model of interest and pool parameter estimates
fit <- with(implist, lmer(JOBSAT ~ I.NEGLEAD +
   G.NEGLEAD + WLOAD + COHES + (1|GRPID)))
testEstimates(fit, var.comp=TRUE)

# test for random slope of leadership style (using D3)
fit2 <- with(implist, lmer(JOBSAT ~ I.NEGLEAD +
   G.NEGLEAD + WLOAD + COHES +
   (1+I.NEGLEAD|GRPID)))
anova(fit, fit2)

# test for contextual effect of leadership style
   (using D1)
context <- "G.NEGLEAD - I.NEGLEAD"
testConstraints(fit, constraint=context)
```

Printed below is the M*plus* syntax that was used for ML estimation of the model of interest.

```
DATA:
file = leadership.dat;

VARIABLE:
names = GRPID JOBSAT COHES NEGLEAD
   WLOAD;
usevariables = JOBSAT COHES NEGLEAD WLOAD
   NEGLEADM;
within = NEGLEAD;
between = COHES NEGLEADM;
cluster = GRPID;
missing = all (-99);

DEFINE:
NEGLEADM = cluster_mean (NEGLEAD);
   ! calculate group means from the observed data
center NEGLEAD (groupmean);   ! group-mean
   centering

ANALYSIS:
type = twolevel;
estimator = ml;

MODEL:
%within%
JOBSAT on NEGLEAD
   WLOAD (1);   ! restrict effect of workload to be
   equal at both levels
NEGLEAD with WLOAD;   ! explanatory variables
   as endogenous, allow covariances
```

```
%between%
JOBSAT on NEGLEADM COHES
  WLOAD (1);  ! restrict effect of workload to be
  equal at both levels
NEGLEADM with COHES;  ! explanatory
  variables as endogenous, allow covariances
NEGLEADM with WLOAD;
COHES with WLOAD;
```

References

Allison, P. D. (2001). *Missing data*. Thousand Oaks, CA: Sage.

Allison, P. D. (2012). Handling missing data by maximum likelihood. In *Proceedings of the SAS Global Forum*. Retrieved from http://support.sas.com/resources/papers/proceedings12/312-2012.pdf

Andridge, R. R. (2011). Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. *Biometrical Journal, 53*, 57–74. http://dx.doi.org/10.1002/bimj.201000140

Asparouhov, T., & Muthén, B. O. (2006). *Constructing covariates in multilevel regression* (Mplus Web Notes No. 11). Retrieved from https://www.statmodel.com/download/webnotes/webnote11.pdf

Asparouhov, T., & Muthén, B. O. (2010). *Multiple imputation with* Mplus (Technical Appendix). Retrieved from http://statmodel.com/download/Imputations7.pdf

Barnard, J., & Rubin, D. B. (1999). Miscellanea: Small-sample degrees of freedom with multiple imputation. *Biometrika, 86*, 948–955. http://dx.doi.org/10.1093/biomet/86.4.948

Bartlett, J. W., Seaman, S. R., White, I. R., & Carpenter, J. R., for the Alzheimer's Disease Neuroimaging Initiative. (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research, 24*, 462–487. http://dx.doi.org/10.1177/0962280214521348

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2016). Lme4: Linear mixed-effects models using 'Eigen' and S4 (Version 1.1–12). [Website.] Retrieved from http://CRAN.R-project.org/package=lme4

Bernaards, C. A., Belin, T. R., & Schafer, J. L. (2007). Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Statistics in Medicine, 26*, 1368–1382. http://dx.doi.org/10.1002/sim.2619

Black, A. C., Harel, O., & Matthews, G. (2013). Techniques for analyzing intensive longitudinal data with missing values. In M. R. Mehl & T. S. Connor (Eds.), *Handbook of research methods for studying daily life* (pp. 339–356). New York, NY: Guilford Press.

Black, A. C., Harel, O., & McCoach, D. B. (2011). Missing data techniques for multilevel data: Implications of model misspecification. *Journal of Applied Statistics, 38*, 1845–1865. http://dx.doi.org/10.1080/02664763.2010.529882

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 3–90). San Francisco, CA: Jossey-Bass.

Bodner, T. E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling, 15*, 651–675. http://dx.doi.org/10.1080/10705510802339072

Carpenter, J. R., & Kenward, M. G. (2013). *Multiple imputation and its application*. Hoboken, NJ: Wiley. http://dx.doi.org/10.1002/9781119942283

Carpenter, J. R., Goldstein, H., & Kenward, M. G. (2011). REALCOM-IMPUTE software for multilevel multiple imputation with mixed response types. *Journal of Statistical Software, 45*, 1–14. http://dx.doi.org/10.18637/jss.v045.i05

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6*, 330–351. http://dx.doi.org/10.1037/1082-989X.6.4.330

Croon, M. A., & van Veldhoven, M. J. P. M. (2007). Predicting group-level outcome variables from variables measured at the individual level: A latent variable multilevel model. *Psychological Methods, 12*, 45–57. http://dx.doi.org/10.1037/1082-989X.12.1.45

Demirtas, H. (2009). Rounding strategies for multiply imputed binary data. *Biometrical Journal, 51*, 677–688. http://dx.doi.org/10.1002/bimj.200900018

Drechsler, J. (2015). Multiple imputation of multilevel missing data—Rigor versus simplicity. *Journal of Educational and Behavioral Statistics, 40*, 69–95. http://dx.doi.org/10.3102/1076998614563393

Enders, C. K. (2008). A note on the use of missing auxiliary variables in full information maximum likelihood-based structural equation models. *Structural Equation Modeling, 15*, 434–448. http://dx.doi.org/10.1080/10705510802154307

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

Enders, C. K. (2011). Missing not at random models for latent growth curve analyses. *Psychological Methods, 16*, 1–16. http://dx.doi.org/10.1037/a0022640

Enders, C. K., Baraldi, A. N., & Cham, H. (2014). Estimating interaction effects with incomplete predictor variables. *Psychological Methods, 19*, 39–55. http://dx.doi.org/10.1037/a0035314

Enders, C. K., Mistler, S. A., & Keller, B. T. (2016). Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. *Psychological Methods*, *21*, 222–240. http://dx.doi.org/10.1037/met0000063

Erler, N. S., Rizopoulos, D., van Rosmalen, J., Jaddoe, V. W. V., Franco, O. H., & Lesaffre, E. M. E. H. (2016). Dealing with missing covariates in epidemiologic studies: A comparison between multiple imputation and a full Bayesian approach. *Statistics in Medicine*, *35*, 2955–2974. http://dx.doi.org/10.1002/sim.6944

Galati, J. C., & Seaton, K. A. (2016). MCAR is not necessary for the complete cases to constitute a simple random subsample of the target sample. *Statistical Methods in Medical Research*, *25*, 1527–1534. http://dx.doi.org/10.1177/0962280213490360

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–472. http://dx.doi.org/10.1214/ss/1177011136

Goldstein, H. (2011). *Multilevel statistical models* (4th ed.). Hoboken, NJ: Wiley.

Goldstein, H., Carpenter, J. R., & Browne, W. J. (2014). Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non-linear terms. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *177*, 553–564. http://dx.doi.org/10.1111/rssa.12022

Goldstein, H., Carpenter, J. R., Kenward, M. G., & Levin, K. A. (2009). Multilevel models with multivariate mixed response types. *Statistical Modelling*, *9*, 173–197. http://dx.doi.org/10.1177/1471082X0800900301

Gottfredson, N. C., Sterba, S. K., & Jackson, K. M. (2017). Explicating the conditions under which multilevel multiple imputation mitigates bias resulting from random coefficient-dependent missing longitudinal data. *Prevention Science*, *18*, 12–19. http://dx.doi.org/10.1007/s11121-016-0735-3

Gottschall, A. C., West, S. G., & Enders, C. K. (2012). A comparison of item-level and scale-level multiple imputation for questionnaire batteries. *Multivariate Behavioral Research*, *47*, 1–25. http://dx.doi.org/10.1080/00273171.2012.640589

Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*, 80–100. http://dx.doi.org/10.1207/S15328007SEM1001_4

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549–576. http://dx.doi.org/10.1146/annurev.psych.58.110405.085530

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, *8*, 206–213. http://dx.doi.org/10.1007/s11121-007-0070-9

Grund, S., Lüdtke, O., & Robitzsch, A. (2016a). Multiple imputation of missing covariate values in multilevel models with random slopes: A cautionary note. *Behavior Research Methods*, *48*, 640–649. http://dx.doi.org/10.3758/s13428-015-0590-3

Grund, S., Lüdtke, O., & Robitzsch, A. (2016b). Multiple imputation of multilevel missing data: An introduction to the R package pan. *SAGE Open*, 6(4), 1–17. http://dx.doi.org/10.1177/2158244016668220

Grund, S., Lüdtke, O., & Robitzsch, A. (2016c). Pooling ANOVA results from multiply imputed datasets: A simulation study. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *12*, 75–88. http://dx.doi.org/10.1027/1614-2241/a000111

Grund, S., Lüdtke, O., & Robitzsch, A. (2018). Multiple imputation of missing data for multilevel models: Simulations and recommendations. *Organizational Research Methods*, *21*, 111–149. http://dx.doi.org/10.1177/1094428117703686

Grund, S., Robitzsch, A., & Lüdtke, O. (2016). Mitml: Tools for multiple imputation in multilevel modeling (Version 0.3–2). Retrieved from http://CRAN.R-project.org/package=mitml

Hill, P. W., & Goldstein, H. (1998). Multilevel modeling of educational data with cross-classification and missing identification for units. *Journal of Educational and Behavioral Statistics*, *23*, 117–128. http://dx.doi.org/10.3102/10769986023002117

Ibrahim, J. G., Chen, M.-H., & Lipsitz, S. R. (2001). Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. *Biometrika*, *88*, 551–564. http://dx.doi.org/10.1093/biomet/88.2.551

Jeličić, H., Phelps, E., & Lerner, R. M. (2009). Use of missing data methods in longitudinal studies: The persistence of bad practices in developmental psychology. *Developmental Psychology*, *45*, 1195–1199. http://dx.doi.org/10.1037/a0015665

Keller, B. T. (2015). *Three-level multiple imputation: A fully conditional specification approach* (Master's thesis). Arizona State University. Retrieved from https://repository.asu.edu/attachments/162109/content/Keller_asu_0010N_15391.pdf

Keller, B. T., & Enders, C. K. (2018). *Blimp user's guide (Version 1.1)*. Retrieved from http://www.appliedmissingdata.com/blimpuserguide-5.pdf

Klein, K. J., Bliese, P. D., Kozlowski, S. W. J., Dansereau, F., Gavin, M. B., Griffin, M. A., . . . Bligh, M. C. (2000). Multilevel analytical techniques: Commonalities, differences, and continuing questions. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research,* *and methods in organizations: Foundations, extensions, and new directions* (pp. 512–553). San Francisco, CA: Jossey-Bass.

Kozlowski, S. W. J., & Klein, K. J. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 3–90). San Francisco, CA: Jossey-Bass.

Kreft, I. G. G., de Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research, 30,* 1–21. http://dx.doi.org/10.1207/s15327906mbr3001_1

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11,* 815–852. http://dx.doi.org/10.1177/1094428106296642

Li, K.-H., Meng, X.-L., Raghunathan, T. E., & Rubin, D. B. (1991). Significance levels from repeated *p*-values with multiply-imputed data. *Statistica Sinica, 1,* 65–92. Retrieved from http://www.stat.sinica.edu.tw/statistica//j1n1/j1n15/j1n15.htm

Li, K.-H., Raghunathan, T. E., & Rubin, D. B. (1991). Large-sample significance levels from multiply imputed data using moment-based statistics and an *F* reference distribution. *Journal of the American Statistical Association, 86,* 1065–1073. http://dx.doi.org/10.1080/01621459.1991.10475152

Licht, C. (2010). *New methods for generating significance levels from multiply-imputed data* (Doctoral dissertation). Universität Bamberg. Retrieved from http://d-nb.info/101104966X/34

Little, R. J. A. (1992). Regression with missing *X*'s: A review. *Journal of the American Statistical Association, 87,* 1227–1237. http://dx.doi.org/10.1080/01621459.1992.10476282

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley. http://dx.doi.org/10.1002/9781119013563

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods, 13,* 203–229. http://dx.doi.org/10.1037/a0012869

Lüdtke, O., Robitzsch, A., & Grund, S. (2017). Multiple imputation of missing data in multilevel designs: A comparison of different strategies. *Psychological Methods, 22,* 141–165. http://dx.doi.org/10.1037/met0000096

Mehta, P. D. (2013). xxM (Version 0.6.0). Retrieved from http://xxm.times.uh.edu

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science, 9,* 538–558. http://dx.doi.org/10.1214/ss/1177010269

Meng, X.-L., & Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika, 79,* 103–111. http://dx.doi.org/10.1093/biomet/79.1.103

Mistler, S. A. (2013). A SAS® macro for computing pooled likelihood ratio tests with multiply imputed data. [Paper 440–2013] In *Proceedings of the SAS Global Forum*. Retrieved from http://support.sas.com/resources/papers/proceedings13/440-2013.pdf

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén. Retrieved from https://www.statmodel.com/download/usersguide/Mplus%20user%20guide%20Ver_7_r3_web.pdf

Newman, D. A. (2003). Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques. *Organizational Research Methods, 6,* 328–362. http://dx.doi.org/10.1177/1094428103254673

Newman, D. A. (2014). Missing data: Five practical guidelines. *Organizational Research Methods, 17,* 372–411. http://dx.doi.org/10.1177/1094428114548590

Nicholson, J. S., Deboeck, P. R., & Howard, W. (2017). Attrition in developmental psychology: A review of modern missing data reporting and practices. *International Journal of Behavioral Development, 41,* 143–153. http://dx.doi.org/10.1177/0165025415618275

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research, 74,* 525–556. http://dx.doi.org/10.3102/00346543074004525

Pinheiro, J., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York, NY: Springer.

Quartagno, M., & Carpenter, J. R. (2016). Jomo: Multilevel joint modelling multiple imputation (Version 2.3–1). Retrieved from http://CRAN.R-project.org/package=jomo

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika, 69,* 167–190. http://dx.doi.org/10.1007/BF02295939

Raghunathan, T. E. (2015). *Missing data analysis in practice*. Boca Raton, FL: CRC Press.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Reiter, J. P., & Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the*

American Statistical Association, *102*, 1462–1471. http://dx.doi.org/10.1198/016214507000000932

Royston, P. (2005). Multiple imputation of missing values: Update. *The Stata Journal*, *5*, 188–201. Retrieved from https://www.stata-journal.com/sjpdf.html?articlenum=st0067_1

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592. http://dx.doi.org/10.1093/biomet/63.3.581

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Hoboken, NJ: Wiley. http://dx.doi.org/10.1002/9780470316696

Särndal, C.-E., Swensson, B., & Wretman, J. (2003). *Model assisted survey sampling*. New York, NY: Springer.

Savalei, V., & Bentler, P. M. (2009). A two-stage approach to missing data: Theory and application to auxiliary variables. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*, 477–497. http://dx.doi.org/10.1080/10705510903008238

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Boca Raton, FL: CRC Press. http://dx.doi.org/10.1201/9781439821862

Schafer, J. L. (2003). Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica*, *57*, 19–35. http://dx.doi.org/10.1111/1467-9574.00218

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177. http://dx.doi.org/10.1037/1082-989X.7.2.147

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*, *33*, 545–571. http://dx.doi.org/10.1207/s15327906mbr3304_5

Schafer, J. L., & Yucel, R. M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics*, *11*, 437–457. http://dx.doi.org/10.1198/106186002760180608

Schafer, J. L., & Zhao, J. H. (2016). Pan: Multiple imputation for multivariate panel or clustered data (Version 1.4). Retrieved from http://CRAN.R-project.org/package=pan

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

Stubbendick, A. L., & Ibrahim, J. G. (2003). Maximum likelihood methods for nonignorable missing responses and covariates in random effects models. *Biometrics*, *59*, 1140–1150. http://dx.doi.org/10.1111/j.0006-341X.2003.00131.x

Taljaard, M., Donner, A., & Klar, N. (2008). Imputation strategies for missing continuous outcomes in cluster randomized trials. *Biometrical Journal*, *50*, 329–345. http://dx.doi.org/10.1002/bimj.200710423

van Buuren, S. (2011). Multiple imputation of multilevel data. In J. J. Hox (Ed.), *Handbook of advanced multilevel analysis* (pp. 173–196). New York, NY: Routledge.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*, 1–67. http://dx.doi.org/10.18637/jss.v045.i03

van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, *76*, 1049–1064. http://dx.doi.org/10.1080/10629360600810434

Vermunt, J. K., & Magidson, J. (2016). Upgrade Manual for Latent GOLD 5.1. Belmont, MA: Statistical Innovations Inc. Retrieved from http://www.statisticalinnovations.com/wp-content/uploads/UpgradeManual5.1.pdf

von Hippel, P. T. (2009). How to impute interactions, squares, and other transformed variables. *Sociological Methodology*, *39*, 265–291. http://dx.doi.org/10.1111/j.1467-9531.2009.01215.x

Wu, L. (2010). *Mixed effects models for complex data*. Boca Raton, FL: CRC Press.

Wu, W., Jia, F., & Enders, C. (2015). A comparison of imputation strategies for ordinal missing data on Likert scale variables. *Multivariate Behavioral Research*, *50*, 484–503. http://dx.doi.org/10.1080/00273171.2015.1022644

Wu, W., West, S. G., & Taylor, A. B. (2009). Evaluating model fit for growth curve models: Integration of fit indices from SEM and MLM frameworks. *Psychological Methods*, *14*, 183–201. http://dx.doi.org/10.1037/a0015858

Yuan, K.-H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology*, *30*, 165–200. http://dx.doi.org/10.1111/0081-1750.00078

Yuan, K.-H., Yang-Wallentin, F., & Bentler, P. M. (2012). ML versus MI for missing data with violation of distribution conditions. *Sociological Methods & Research*, *41*, 598–629. http://dx.doi.org/10.1177/0049124112460373

Yucel, R. M. (2008). Multiple imputation inference for multivariate multilevel continuous data with ignorable non-response. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *366*, 2389–2403. http://dx.doi.org/10.1098/rsta.2008.0038

Yucel, R. M., Ding, H., Uludag, A. K., & Tomaskovic-Devey, D. (2008). Multiple imputation in multiple classification and multiple-membership structures. In *Proceedings of the Section on Bayesian Statistical Science of the American Statistical Association*. 4006–4013. Retrieved from https://ww2.amstat.org/sections/srms/Proceedings/y2008/Files/302707.pdf