Psychological Test and Assessment Modeling, Volume 63, 2021 (3), 432-455

Estimation of partially observed non-linear terms in a multilevel model
An evaluation of the robustness of ad hoc and state-of-the-art missing data methods

Kristian Kleinke[1]

Abstract

Multiple imputation (MI) of outcome variables and linear predictors has received the most attention so far in the MI literature. Research regarding imputation of incomplete non-linear terms (like interaction terms or quadratic or even higher-order relationships) on the other hand is still scarce. The present paper examined two ad hoc MI strategies and two more theoretically sound MI solutions (i.e. substantive model compatible MI) regarding bias in statistical inferences in a random intercept model that includes an interaction term and a quadratic term by means of Monte Carlo simulation. The distribution of predictor variables was either normal, heavy-tailed or skewed. Results show that only if the imputation model is fully compatible to the subsequent analysis model (and to the original data generating process), i.e. only when the imputation model includes non-linear terms as well as information regarding cluster membership, then point estimates and standard errors are unbiased. Currently available MI methods therefore need to be adjusted for situations, where distributional assumptions are violated to some extent.

Keywords: missing data · multiple imputation · multilevel model · incomplete predictors

[1] *Correspondence concerning this article should be addressed to:* Kristian Kleinke, University of Siegen, Institute of Psychology, Adolf-Reichwein-Str. 2a, D-57068 Siegen; email: Kristian.Kleinke@uni-siegen.de

Introduction

Multiple Imputation (MI) of outcome variables and linear predictors both at level-1 and level-2 has received the most attention so far in the MI literature about incomplete multilevel data (e.g. Enders, Mistler, & Keller, 2016; Grund, Lüdtke, & Robitzsch, 2016; Grund, Lüdtke, & Robitzsch, 2018a, 2018b; Lüdtke, Robitzsch, & Grund, 2017; Kleinke & Reinecke, 2015; Kleinke, Stemmler, Reinecke, & Lösel, 2011). Research regarding how to handle incomplete predictors—when these predictors are used to form non-linear associations like quadratic terms or interaction terms on the other hand is still scarce. Little is yet known about the properties of the respective solutions regarding the analysis of incomplete multilevel data.

Yet in applied research, both the inclusion of higher-order relationships (e.g. to model u-shaped or inverted u-shaped relationships with the dependent variable) or interaction terms (to test for moderator effects) are highly relevant.

Currently there are various options available to handle these kinds of situations—both ad hoc solutions, and also more theoretically sound methods. The present paper focusses on 4 methods: (1) passive imputation (PI), which does not impute variables that are functions of other variables, but computes their scores using the imputed values of these other variables, (2) the 'just another variable' method (JAV), which simply imputes a variable that is a function of other variables as just another variable (von Hippel, 2009), and finally, what has come to be known as 'substantive model compatible multiple imputation' both under the (3) joint modelling MI framework (e.g. R package jomo, Quartagno & Carpenter, 2020) or (4) under the conditional modelling MI framework (e.g. R package smcfcs, Bartlett & Keogh, 2020).

All methods currently have advantages and disadvantages. The obvious advantage of the two ad hoc methods (passive imputation and the just another variable method) is their ease of use. These strategies could for example be applied using standard MI software in R (e.g. mice, van Buuren, 2012; van Buuren & Groothuis-Oudshoorn, 2011) and many other packages. Research by von Hippel (2009) and Seaman, Bartlett, and White (2012), however, already suggests that both solutions could produce biased statistical inferences: As will be discussed more thoroughly in the next section, passive imputation usually leads to a misspecified imputation model, and—depending on the severity—biased inferences. The 'just another variable' method on the other hand only appears to work well when the data are missing completely at random (MCAR) in the sense of Rubin (1976)—a restrictive assumption that is often violated in real life situations. Note furthermore that when data are MCAR, imputation is usually not required. Apart from preventing potential loss of statistical power due to exclusion of cases with missing values, application of MI does not have any benefits over methods that are based on the remaining observed cases.

In contrast to the simple ad hoc methods, substantive model compatible MI approaches—as the name suggests—create imputations of incomplete data using a model that is compatible to the subsequent analysis model and incorporates variables that are functions of other variables explicitly into the model. Solutions are available for both popular MI frameworks—conditional modelling and joint modelling (see the next section for details): substantive model compatible fully conditional specification (Bartlett, Seaman, White, & Carpenter, 2015), as it is for example implemented in R-package smcfcs (Bartlett & Keogh, 2020), and substantive model compatible MI based on the joint modelling MI framework (Goldstein, Carpenter, & Browne, 2014), as it is for example implemented in R package jomo (Quartagno & Carpenter, 2020).

One drawback of the smcfcs-Package at the moment however is that it currently only includes functions for multivariate data, but not for clustered data. One aim of the present paper is to evaluate, if the functions could also be applied in a multilevel scenario, where cluster effects are not too large.

The jomo package in contrast does include functions for multilevel data, however, at the moment, only very few Monte Carlo Simulations have systematically evaluated the package (e.g. Grund et al., 2018a), and usually not regarding the imputation of non-linear terms. Also, at the moment there is no great abundance of substantive models that are supported (a two-level “normal” model, and a two-level logit model). Usually, data analysts do not really care about the distribution (or the ‘normality’) of predictor variables, since standard regression type models do not make any assumptions regarding the distribution of these variables. In the context of multiple imputation of partially observed predictor variables, however, these predictors become the outcome variables in the imputation models, and we need to think about the conditional distribution of these variables given the other variables in the imputation model. So, the ‘normality’ of predictor variables suddenly does become an issue—when no other imputation methods and models are available. One focus of this paper therefore also is to explore, to what extent the currently available methods for partially observed predictors can cope with predictor variables that are not normally distributed, but for example more heavy-tailed or skewed.

The present paper thus seeks to broaden the knowledge regarding the applicability of the currently available approaches, as well as to identify avenues for future research and software development.

The paper is structured as follows. First, I give a brief introduction to multiple imputation with a focus on multilevel data, and discuss the aforementioned solutions to impute non-linear terms in multilevel models. I then present three Monte Carlo Simulations, in which I evaluated these solutions in scenarios, where distributional assumptions were either met, or mildly or even severely violated, to see how robust these methods and models are. The paper ends with a discussion of the findings and by outlining avenues for future research and software development.

Theoretical Background

Multiple Imputation is one of the standard methods to address the missing data problem both in multivariate data sets, as well as in clustered data structures (for an introduction and comprehensive overview, see for example Kleinke, Reinecke, Salfrán, & Spiess, 2020). MI can make use of all available information in the data file to predict missing information, and is especially worthwhile in situations where missingness depends on observed variables in the data file that need not even be of interest in the data analysis model later on.

Application of MI is a three-step-process: In the first step each missing value is replaced multiple times based on some statistical model (that should be at least as detailed as the subsequent analysis model, i.e. should include all the variables and associations that are of interest to the data analyst. In the second stage, the m completed date files are analyzed separately using any confirmatory complete-data method. Thirdly, the m sets of model results (i.e. point estimates and their standard errors) are combined into an overall set of results using standard formula (Barnard & Rubin, 1999; Rubin, 1987). The combined parameter estimate is simply the average across the imputations. The MI standard error estimate combines variation between and within the m imputed data sets and reflects the extra estimation uncertainty due to missing data.

Two popular frameworks for creating the m sets of imputations are joint modelling (JM) and conditional modelling (CM). The first framework has a sound theoretical foundation in Bayesian statistics (e.g. Schafer, 1997) and requires specification of a joint model for all incomplete variables to be imputed. However, in practice it is often difficult (or even impossible) to find a joint model that fits the problem at hand well. The second framework tries to overcome this problem. Conditional modelling is much more pragmatic and only implicitly assumes an underlying joint distribution. CM (e.g. van Buuren, 2012) allows to specify separate imputation models for each incompletely observed variable in the data file. Although conditional modelling lacks the sound theoretical foundation, various Monte Carlo Simulation have shown that the method nevertheless works well in many practically relevant scenarios (for a more in-depth discussion, see Kleinke et al., 2020, Chapter 4).

Regardless of the MI framework that is being used, to obtain unbiased statistical inferences, some requirements have to be met: Missing data are assumed to be missing at random (MAR) in the sense of Rubin (1976), which—simplistically speaking—means that missing information can be predicted by observed information in the data file, and that missingness does not additionally depend on unobserved information. Furthermore, the prediction model for the incomplete information needs to be at least as detailed as the subsequent analysis model, i.e. must include all the variables of the model of scientific interest, including all relevant higher order relationships, interactions, as well as information regarding cluster membership. If the imputation model fails to include any of this information, the respective effect will be biased towards zero (e.g. Drechsler, 2015; Schafer, 1997) [2]. Ideally, we thus want both informative data and a ‘good’ imputation model that includes all relevant observed information to predict missing information (see also Collins, Schafer, & Kam, 2001). Too ‘large’ imputation models are usually not problematic. In fact, MI was originally designed for large public use data files, and especially for situations, where ‘the imputer’ knows more than ‘the analyst’ (i.e. where the data provider could use observed information for the prediction of incomplete information that cannot be disclosed to the analyst). Nowadays, MI has become a standard procedure also for much smaller data sets, as they are common in the field of psychology, and MI has relevance also in the context of psychological testing and assessment modelling (see for example this journal’s special topic on missing values, here especially the papers by Aßmann, Gaasch, Pohl, & Carstensen, 2015; Köhler, Pohl, & Carstensen, 2015; Vidotto, Vermunt, & Kaptein, 2015; Vink, Lazendic, & van Buuren, 2015).

Multiple imputation of multilevel data

When the model of scientific interest is a multilevel model, one important choice that needs to be made is whether to apply single-level or multilevel imputation techniques. Both variants currently have advantages and disadvantages—the most important problem currently being the sparseness of adequate MI software for multilevel data imputation. Most statistical software packages currently only include basic MI implementations like Schafer’s (1997) algorithms for creating imputations based on the joint multivariate normal model or regression based approaches like k nearest neighbor imputation based on predictive mean matching (Little, 1988). Multilevel imputation models are supported only by some packages. Additionally, regarding the imputation of partially observed predictor variables that are functions of other variables, only very few packages have implemented this functionality, and there is not a huge variety

[2] Note that there are exceptions to this general rule. In randomized controlled trials (RCT), European Medicines Agency (EMA) guidelines on missing data recommend conservative methods that do not overestimate treatment effects (Committee for Medicinal Products for Human Use, 2010). When the imputation model is correctly specified, a proper imputation method in the sense of Rubin (1987, 1996) is used, and when all modeling assumptions are at least approximately met, estimated treatment effects could be expected to be unbiased. In fact, since uncertainty about the parameter is reflected by the between imputation variance component (and consequently also in standard errors and confidence intervals), MI based on Rubin’s theory can be per se regarded as conservative meeting the EMA guidelines. In practice, however, potentially improper MI methods such as predictive mean matching are often used as a remedy against bias introduced by violations of distributional assumptions (regarding the properness of pmm, see Gaffert, Meinfelder, & Bosch, 2018). Depending on the respective scenario, this works more or less well, and some bias can be expected (e.g. K. Kleinke, 2017, 2018). To avoid overestimations of treatment effects, Twisk et al. (2020) propose a strategy, which they refer to as “selective imputation”: here missing values of participants from the intervention group that have never received the treatment (due to drop out after the pre measurement timepoint) are imputed as if they were control patients—thus producing a conservative estimate.

of models to choose from. So applied researchers have to make some trade-offs, which I would like to discuss in this section.

Due to the sparseness of available MI software that allows to create imputations based on multilevel models, in practice, the multilevel structure of the data is often ignored at the imputation stage (see the discussions in Graham, 2009 and Drechsler, 2015). This of course oversimplifies the imputation model. Usually, the obtained fixed effects estimates, however, are nevertheless acceptable, and bias is usually to be expected only regarding the ‘random’ part of the model. Since random effects are not part of the imputation model, subsequent random effects estimates of the substantive model will be biased towards zero (see also the simulation in van Buuren, 2011). Since the magnitude of bias, inter alia, depends on the number of missing values to be imputed, and the size of the random effects variances (see also the discussion in Drechsler, 2015, and the references therein), ignoring the clustered structure of the data might be an acceptable strategy for applied researchers, when random effects variances are not too large, when correct estimation of random effects variances is not of scientific importance, or when the missing data problem is rather minor. Kleinke et al. (2011) have for example compared this strategy against a two-level imputation approach in a systematic simulation based on empirical clustered data with a missing data percentage of only up to 20%, and found that the obtained inferences were widely unbiased and comparable to the ones obtained by the two-level imputation method.

Of course, generally speaking, the theoretically more appropriate strategy would be to fill in incomplete clustered data using an MI method that explicitly allows for clustering. Multilevel MI methods are available for both MI frameworks—joint and conditional modelling, however, currently few jont modelling programms support substantive model compatible MI, like jomo (Quartagno & Carpenter, 2020) or REALCOM-IMPUTE (Carpenter, Goldstein, & Kenward, 2011). REALCOM-IMPUTE (Carpenter et al., 2011) builds upon the pan algorithm—one of the earliest joint modelling solutions for multilevel data, proposed by Schafer and Yucel (2002). pan generates multiple imputations of continuous incomplete variables from a multivariate linear mixed effects model. One major disadvantage of pan, however, was that it did not allow for incomplete predictors. REALCOM-IMPUTE on the other hand now allows incomplete predictors at all levels. The pan method generally works fine when all model assumptions (like normality of level-1 errors, normality of random effects, homoscedasticity) are met, the imputation model is correctly specified, and the data are missing at random (see for example the simulation in Enders et al., 2016). An R adaptation of REALCOM-IMPUTE is package jomo (Quartagno & Carpenter, 2020). While multilevel imputation functions are also available in the conditional modelling framework—R package mice (van Buuren & Groothuis-Oudshoorn, 2011) for example includes functions to impute continuous two-level data based on a ‘normal’ linear mixed effects model, and further packages provide additional add-on functions for example for binary or count data (Kleinke & Reinecke, 2019; Robitzsch, Grund, & Henke, 2017)—neither of these functions currently support substantive model compatible MI.

It needs to be noted that most of the above packages are furthermore based on rather restrictive parametric models. If all modelling assumptions are (at least approximately) met, and the models are correctly specified, then the imputation functions are expected to allow valid inferences. Unfortunately, due to scarce simulation results, it is not yet clear, if or to what extent these methods are robust against violations of their distributional assumptions.

Multiple imputation of variables that are functions of other variables

In addition to the choice whether to use single-level or multi-level imputation models, when it comes to incomplete predictors that are functions of other variables, applied researchers have further choices and trade-offs to make: to apply a relatively simple ad-hoc strategy, or to apply a more theoretically sound model-based solution (that at the moment does not support many different models).

Passive imputation

is a straightforward strategy that can easily be applied using standard MI software. Passive imputation omits all variables that are functions of other variables from the imputation model and forms them passively from the imputed values afterwards. When the substantive model of scientific interest includes an interaction term of variables x_1 and x_2, we would only impute missing values in x_1 and x_2, and then compute the interaction term x_3 by taking the product of the imputed values of x_1 and x_2: $x_{3i} = x_{1i}x_{2i}$. When, for example, the imputed value of x_1 for a certain case i is 4, and the imputed value of x_2 for that case is 3 then the interaction x_1x_2 would simply be computed as $4 \cdot 3 = 12$. Passive imputation ensures that all values in x_1, x_2, and x_3 are consistent for all cases in the data file, i.e. that x_{3i} is always $x_{1i}x_{2i}$ for all cases $i = 1,2, \dots, n$. The main goal of passive imputation thus is to maintain the consistency between the original variable(s) and the transformed or combined variables.

The main problem of passive imputation, however, is that it usually leads to a misspecified imputation model (see the discussion in Seaman et al., 2012): Passive imputation makes an imputation method ‘improper’, if the transformed variable has an extra effect on the dependent variable in the model of applied-scientific interest. To illustrate this, let us assume our substantive model of interest is a standard linear regression model,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon, \quad E(\epsilon) = 0,\ \text{var}(\epsilon) = \sigma^2. \tag{1}$$

in which dependent variable y is regressed on predictor x_1 and its quadratic term $x_2 = x_1^2$. Let us furthermore assume that y is completely observed, and x_1 is only partially observed. Adopting a regression based MI approach, the imputation model for missing values in x_1 could be

$$x_1 = \gamma_0 + \gamma_1 y + \nu. \tag{2}$$

Following the passive imputation idea, we would then obtain x_2 by squaring the imputed x_1 values. The problem, however, is that (2) is the 'correct' model only when the 'true' coefficient of β_2 in (1) is zero (Seaman et al., 2012) and $(y,\ x)$ is is normally distributed. If $\beta_2 \neq 0$, bias is to be expected, because imputed values of x_1 only reflect the linear relationship between x_1 and y.

The just another variable method

in contrast to passive imputation does include variables that are functions of other variables into the imputation model and imputes them—as the name says—as '*just another variable*'. In the example above, we would impute both x_1, x_2, as well as the interaction term x_3. This however usually leads to situations where $x_{3i} \neq x_{1i}x_{2i}$, which might not be problematic since multiple imputation does not aim to make plausible predictions on the person level, but instead aims at making valid statistical inferences. Unfortunately, Seaman et al. (2012) have found, that the just another variable method only appears to work sufficiently well when the data are MCAR—an assumption that is often violated in practice. When values are MAR then the missing data mechanism may selectively thin out certain regions of the sample space. If this occurs systematically in a region were the true relationship deviates considerably from the relationship assumed in the imputation model, then bias is possible.

One aim of the present simulation is to further corroborate these findings.

Substantive model compatible fully conditional specification

is a solution proposed by (Bartlett et al., 2015), which is a modification of the traditional conditional modeling mice framework, so that covariates could be imputed from the 'appropriate' model. Appropriate in this context means that the imputation model is compatible to the (assumed) data generating model. Unfortunately, generalizations of this method have not yet been made with regards to multilevel models and the restrictions and shortcomings that were discussed in the previous section apply, when using this method for clustered data structures. This means: imputations of model covariates could (at the moment) only be created under the restrictive assumption that no cluster effects have to be assumed for the covariates in question. As discussed above, although the clustered structure is ignored during the imputation stage, it could still be expected that the method produces widely unbiased point estimates in the fixed part of the model, while inferences regarding the random part of the model could be biased. One aim of the present paper is to elucidate, if smcfcs in its current version could be safely applied to impute incomplete covariates in a multilevel scenario. Since bias is known to depend (inter alia) both on the missing data fraction and the magnitude of cluster effects, I expect the method to work reasonably well, when not not too many values are missing, and random effects variances are rather small.

Jomo.

An alternative to smcfcs is the Bayesian joint modelling procedure proposed by Goldstein et al. (2014), which has been implemented for example in R package jomo. The method ensures compatibility of the imputation model and the substantive model (i.e. the data analyst's model) by defining the imputation model as the conditional distribution of the outcome variables given the covariates multiplied by the joint distribution of the covariates. The method assumes a joint multivariate model for all variables in the imputation model. jomo at the moment supports a 'normal' linear mixed effects model and a binary generalized linear mixed effects model. The explicit advantage over smcfcs is that the multilevel structure of the data could be considered during the imputation stage. I would thus expect that when the data generating model is a multilevel model (and all modelling assumptions are met), that jomo will produce the best results in terms of bias in parameter estimates and standard errors in comparison to the other methods.

Little is yet known regarding the robustness towards violations of modelling assumptions. One aim of the present paper is to broaden the knowledge in this regard. The simulations shall elucidate to what extent the discussed approaches are robust regarding violations of these assumptions. In addition to a situation where all parametric assumptions are met (simulation 1), simulations 2 and 3 will look at scenarios where predictors have a symmetric but more heavy-tailed distribution in comparison to the normal distribution, or where predictors are skewed.

Method

Overview

To evaluate, how the different MI approaches fare in terms of obtaining unbiased statistical inferences in a multilevel scenario, I simulated data based on a model that is highly similar to the one used in Goldstein et al. (2014, Sect. 8). The simulated data in this scenario could for example represent a total of 4000 students from 50 schools in a large city. From each school, 80 students are selected into the sample.

The data generating model was a random intercept model (e.g. Bryk & Raudenbush, 1992), i.e. we assume school-level differences regarding the intercepts:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{1ij} x_{2ij} + \beta_4 x_{2ij}^2 + u_j + e_{ij}, \qquad (3)$$

with $\beta_0 = 0$, $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0.5$, $u_j \sim \mathcal{N}(0, \sigma_u^2 = 0.1)$, and $e_{ij} \sim \mathcal{N}(0, \sigma_e^2 = 0.5)$.

In (3) coefficients β denote the fixed effects—the average parameter estimates across all participants. x_1 and x_2 might be test scores from a psychological test. u_j denote cluster-specific deviations from the intercept of cluster j, and e_{ij} are level-1 residuals of individual i. The intercept variance σ_u^2 was set as 0.1, the level-1 resiual variance as $\sigma_e^2 = 0.5$. β_3 reflects the interaction of x_1 and x_2, β_4 the non-linear relationship (here: quadratic) between x_2 and the outcome variable y.

In the first simulation, the two level-1 predictors x_1 and x_2 were drawn from a multivariate normal distribution with mean 0 and covariance 0.5

$$(x_1, x_2) \sim \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1{,}0.5\\0.5{,}1\end{bmatrix}\right). \tag{4}$$

In the second simulation, x_1 and x_2 were drawn from a multivariate t distribution with 3 degrees of freedom

$$(x_1, x_2) \sim t_3\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1{,}0.5\\0.5{,}1\end{bmatrix}\right), \tag{5}$$

making the shape more heavy-tailed in comparison to the first simulation.

In the third simulation, predictors x_1 and x_2 were drawn from a χ^2-distribution with 3 degrees of freedom. Data were generated in the following way. Firstly, two normal variables were generated like in simulation 1. For these values I obtained the value of the standard normal distribution function. The respective quantiles to these values of the χ^2-distribution with 3 degrees of freedom were then taken as the respective x_1 and x_2 values. In comparison to the first simulation this lead to a positively skewed distribution.

I then introduced 20% missing completely at random (MCAR) missingness (cf. Rubin, 1976) in both x_1 and x_2.[3] MCAR means that the missingness mechanism is a completely random process that neither depends on observed information in the data set nor on unobserved information. MCAR missingness was chosen so that the respective MI methods could be evaluated not only in comparison to the set population quantities, but additionally also against complete case analysis, which is likely to produce unbiased inferences, when the MCAR-assumption holds, and which might be a better option than MI in practice, when MCAR more or less holds.

Each scenario (normal, t, χ^2 distributed predictors) was replicated 1000 times.

[3] Note that the focus of the present simulations were problems regarding incomplete predictors, when these predictors are used to form non-linear terms, and also problems regarding distributional assumptions of these imputation models. To this end, I kept the simulation set up simple focussing on these aspects alone. In empirical applications, values would usually also be missing in the outcome variable.

Missing Data Handling

Just another variable method

In the first simulation (normally distributed predictors), the JAV method was applied within the conditional modelling MI framework using the mice package in R (van Buuren & Groothuis-Oudshoorn, 2011). JAV means that variables that are functions of other variables are imputed just as any another incomplete variable. Therefore, in addition to y, x_1, and x_2, two further variables were computed

$$x_3 = x_1 x_2$$
$$x_4 = x_2^2$$

and included into the imputation model. Missing values in x_1–x_4 were imputed based on a Bayesian 'normal' linear regression model (i.e. method "norm" in mice). 100 imputations were created respectively. The number of iterations for mice's Gibbs sampler was set to 20. Convergence diagnostics were performed as outlined in (van Buuren & Groothuis-Oudshoorn, 2011) and van Buuren (2012).

In the second (heavy-tailed predictors) and third simulation (skewed predictors), I used the same imputation strategy, here however, I used predictive mean matching (method "pmm" in mice) rather than 'normal' Bayesian linear regression to fill-in the missing values. Predictive mean matching (pmm) is a k nearest neighbour imputation method: Based on the values predicted by a 'normal' linear regression model, an observed donor case is sampled from a pool of $k = 5$ cases, whose predicted means are closest to the one for the incomplete case. This cases's observed value is then used to fill-in the missing one. Previous research has shown that pmm produces better results (in terms of unbiased parameter estimates and standard errors) than normal linear regression, when parametric assumptions are mildly to moderately violated (Kleinke, 2017).

Passive imputation

was applied using the mice package in R. The imputation model was a "normal" linear regression model (method "norm") in the first simulation. Predictive mean matching was used in the second and third simulation. Missing values in variables x_3 (the interaction term) and x_4 (the quadratic term) were computed passively from the imputed values of x_1 and x_2 in each cycle of mice's Gibbs sampler. Again, $m = 100$ imputations were generated.

Conditional modelling substantive model compatible MI

was applied using package smcfcs (Bartlett & Keogh, 2020). Note that robust imputation methods are currently not supported. Thus for all simulations, a normal linear regression model was applied for x_1 and x_2. The non-linear terms were imputed passively under the appropriate model. Note furthermore, that random effects (multilevel) models are currently also not supported. Again, $m = 100$ imputations were generated.

Joint modelling substantive model compatible MI

was applied using function jomo.lmer from package jomo (Quartagno & Carpenter, 2020). Here (at least in simulation 1), the imputation model was fully compatible to the data generating model and to the subsequent analysis model. Note that jomo at the moment supports normal and binary outcome variables. Violations of distributional assumptions regarding incomplete predictors like the ones from simulations 2 and 3 at the moment cannot explicitly be modelled. Again, $m = 100$ imputations were created, using the package's default settings, i.e. the first set of imputations is drawn after a burn-in period of 1000 cycles of the Markov Chain Monte Carlo algorithm, and the sets of imputations are drawn after another 1000 cycles each to ensure stochastic independence of the imputations.

Substantive model and Monte Carlo quality criteria

The completed data files were then analyzed by a random intercept model including variables x_1, x_2, their interaction x_1x_2 and the quadratic term x_2^2. Models were fitted and combined using functions from R package mitml (Grund, Robitzsch, & Luedtke, 2021), which calls lme4 (Bates, Mächler, Bolker, & Walker, 2015) to repeatedly estimate the random intercept model across the m sets of imputations. The standard large sample formula were used for obtaining multiple imputation inferences (Rubin, 1987).

To evaluate the quality of the respective missing data solutions, I report bias in parameter estimates and 95% confidence interval coverage rates.

Bias is defined as the difference between the defined 'true' population parameter and its average estimate across the 1000 Monte Carlo replications, and reflects the accuracy of point estimates. Bias obviously shall be close to zero. Negative bias indicates that the average estimate is larger than the true parameter. Following Forero and Maydeu-Olivares (2009), bias is deemed significantly large if it exceeds more than 10% of the true parameter (boldface type in Tables 1–3).

95% confidence interval coverage of a certain parameter is defined as the percentage of 95% confidence intervals (across the 1000 Monte Carlo replications) that include the 'true' population parameter. Obviously, this percentage shall be close to 95% (cf. Schafer & Graham, 2002). Undercoverage, i.e. rates below 90% could either indicate too large biases (so that the interval is shifted too far to the right or to the left to include the true parameter), or too small standard erros (so that the interval is not wide enough to include the true parameter. On the other hand, very large coverage rates close to 1 could indicate too large standard errors and consequently too wide confidence intervals (which increases the risk of type II errors).

Taken together, bias and coverage rates give a good impression of the quality of the respective approach in a given scenario.

Results

Simulation 1

Results of the first simulation are summarized in Table 1. Firstly, to check that the data simulation process worked well, I obtained the average parameter estimates across the set of 1000 complete data sets (i.e. before any missing data were introduced). Rounded to three decimal places the complete data estimates were -0.001 for the intercept term, 0.500 respectively for $\beta_1 - \beta_4$, and 0.100 for the intercept variance. The very small deviation (regarding the intercept) from the defined population parameter reflects the usual sampling error.

Table 1:

Performance of various strategies to impute multivariate normal covariates in a random intercept model

	β_0	β_1	β_2	β_3	β_4	σ_u^2
Q	0	0.500	0.500	0.500	0.500	0.100
			Bias			
OC	0.001	0	0	0	0	0
PI	−0.130	**0.055**	0.019	**0.123**	**0.070**	0.007
JAV	−0.004	0.005	0	0.009	0	**0.037**
smcfcs	−0.005	0.006	0	0.005	0.002	**0.026**
jomo	0.001	0.001	−0.001	−0.002	0.003	0
			Coverage Rate in %			
OC	96	94	94	95	96	–
PI	**25**	**30**	92	**0**	**3**	–
JAV	**89**	94	95	91	95	–
smcfcs	91	94	95	95	97	–
jomo	95	96	94	94	94	–

Note. Biases were rounded to three decimal places. Q is the 'true' population parameter. OC refers to the estimates based on the available observed cases. PI is passive imputation, JAV the just another variable approach. smcfcs is substantive model compatible fully conditional specification, jomo is substantive model compatible joint modelling. β are the coefficients in the fixed part of the model, σ_u^2 denotes the intercept variance. Use of boldface type in the top part of the table indicates large absolute biases that were more than 10% the size of the true parameter. Use of boldface type in the bottom part indicates seriously low coverage, i.e. below 90%. Standard errors and confidence intervals for the random part of the model were not computed.

Results of the various missing data methods are displayed in Table 1. As expected, model estimates based on the remaining observed cases (OC in Table 1) yielded unbiased statistical inferences. Biases were in fact virtually zero and coverage rates ranged between 94% and 96%, which indicates that obtained confidence intervals were adequately large. Secondly, passive imputation yielded overall the largest absolute biases and the lowest coverage rates with an average of only 30% across all parameters. The worst parameters here were the two non-linear terms, the interaction term between the two predictors with a coverage rate of virtually zero, and the quadratic term also with a 95% confidence interval coverage of very close to zero.

The just another variable method on the other hand yielded acceptable inferences for most parameters. With the exception of the intercept variance, biases were typically small, and with the exception of the intercept, all coverage rates lay within an acceptable range. The average coverage rate across all parameters was 93%. Bias regarding the intercept variance was 0.037, which in relative terms is 37% of the true parameter. Following Forero and Maydeu-Olivares (2009), relative biases (defined as absolute bias devided by the value of the true parameter) of larger than 10% can be regarded as inadequately large.

The conditional modelling based substantive model compatible MI approach also yielded low biases for the fixed part of the model. Across all parameters the average coverage rate was 95%. As was the case for the JAV method, the estimate of the intercept variance was off. In relative terms, bias here was 26% of the true parameter, which is unacceptably large. Note that JAV, as well as the substantive model compatible approach did not consider the multilevel structure of the data during the imputation stage and some bias regarding the random part of the model was to be expected.

Finally, compared to all other imputation methods, jomo yielded the overall 'best' results in terms of low biases and adequate 95% confidence intervall coverage. Mean absolute bias across all parameters was 0.001, relative biases (not shown in Table 1) were all below 10%, and coverage ranged between 94% and 96%. The imputation model used by jomo was the only model that was fully compatible to the subsequent analysis model.

Simulation 2

Results of the second simulation are summarized in Table 2. Again, analyses of the complete data sets (before any missing data were introduced) revealed that the data generating process worked well. Since (multilevel) regression models do not make any assumptions regarding the distribution of predictors, and since the data in these predictors were MCAR, analysis based on the remaining observed cases (case deletion) were unbiased and all coverage rates lay in an acceptable range.

Like in simulation 1, PI did not produce acceptable results. Coverage rates ranged between 66% and 88% with an average of 80% across all parameters. Relative biases ranged between 1% and 47% of the true parameter with an average relative bias of 16%. One noteworthy finding is that coverage rates were considerably larger in comparison to simulation 1. Using predictive mean matching (which imputes an actual observed case) seems to buffer some of the shortcomings of passive imputation regarding the adequate imputation of variables that are functions of other variables. Since this buffer effect might be specific to this particular scenario, future research should look into this more thoroughly.

Adopting the JAV method produced an average coverage rate of 78% (range: 70%–84%). Relative biases ranged between 4% and 14%. Largest biases were found for the coefficients of x_1, x_2, and their interaction.

smcfcs also produced seriously low coverage rates of between 54% and 78% with an average coverage of 64% across all parameters. Mean relative bias was 10%. Like in simulation 1, the estimate for the intercept variance was biased. Additionally, unacceptably large biases were found for the parameters of x_1 and the corresponding interaction with x_2.

Again, of all imputation methods, jomo produced the most accurate point estimates. All relative biases were below 10%, and the average relative bias was only 0.21%. Coverage rates ranged between 83% and 92% with an average of 88% across all parameters. It seems that while point estimates were unbiased, estimates of standard errors were generally too low to produce acceptable confidence interval coverage.

Table 2:
Performance of various strategies to impute multivariate t-distributed covariates in a random intercept model

	β_0	β_1	β_2	β_3	β_4	σ_u^2
Q	0	0.500	0.500	0.500	0.500	0.100
			Bias			
OC	0	0	0	0	0	0
PI	−0.139	**0.066**	0.004	**0.083**	−0.008	−0.047
JAV	0.160	−0.058	−0.069	−0.018	−0.062	−0.009
smcfcs	−0.060	**0.059**	−0.010	−0.076	0.032	−0.016
jomo	0.016	0.008	−0.004	−0.003	0.003	0
			Coverage Rates in %			
OC	93	96	95	95	94	–
PI	**79**	**66**	**82**	**83**	**88**	–
JAV	**70**	**83**	**79**	**84**	**75**	–
smcfcs	**78**	**56**	**75**	**58**	**54**	–
jomo	92	**86**	92	**88**	**83**	–

Note. Biases were rounded to three decimal places. Q is the ‘true’ population parameter. OC refers to the estimates based on the available observed cases. PI is passive imputation, JAV the just another variable approach. smcfcs is substantive model compatible fully conditional specification, jomo is substantive model compatible joint modelling. β are the coefficients in the fixed part of the model, σ_u^2 denotes the intercept variance. Use of boldface type in the top part of the table indicates large absolute biases that were more than 10% the size of the true parameter. Use of boldface type in the bottom part indicates seriously low coverage, i.e. below 90%. Standard errors and confidence intervals for the random part of the model were not computed.

Simulation 3

Results of the third simulation are displayed in Table 3. Firstly, like in the previous simulations, analyses based on the remaining observed cases worked well—both in terms of unbiased point estimates as well as unbiased measures of uncertainty.

Passive imputation produced coverage rates between 2% and 59% (average: 34% across all parameters) and biases were large for most parameters. Largest relative biases were 57% and 61% for β_1 and β_2 respectively. Average relative bias was 40% across all parameters.

The JAV approach yielded an average coverage rate of 85% with a range between 73% and 96%. Fixed effects biases were small, and below 10%, with the exception of β_2. The larges relative bias was 52% for the intercept variance.

smcfcs yieled inacceptable large biases and consequently also small coverage rates for some parameters. Large biases were found for the coefficients of x_1 and x_2, and the estimate of the intercept term was abysmal (alsmost twice as large as the bias produced by PI). Also the estimate of the intercept variance was biased (12%).

Again, of all imputation methods, jomo produced the overall most accurate point estimates. Mean relative bias was 1%. Like in simulation 2, while point estimes were widely unbiased, confidence intervals seemed to be too narrow (i.e. the obtained standard errors are too small). Mean coverage rate was 76%, the range was 52% – 88%. While in simulation 1, coverage rates lay within an acceptable range, violations of normality (here: skewed predictor variables) seem to affect the obtained MI standard errors.

Table 3:

Performance of various strategies to impute χ^2-distributed covariates in a random intercept model

	β_0	β_1	β_2	β_3	β_4	σ_u^2
Q	0	0.500	0.500	0.500	0.500	0.100
Bias						
OC	0	0.001	0	0	0	0
PI	0.491	−0.285	−0.304	**0.166**	−0.041	**0.041**
JAV	0.023	−0.029	**0.067**	0.002	−0.011	**0.052**
smcfcs	−0.918	**0.109**	**0.291**	−0.030	−0.014	**0.012**
jomo	−0.068	0.008	0.028	−0.002	−0.002	0
Coverage rates in %						
OC	95	94	95	92	94	–
PI	**59**	**29**	**41**	**2**	**39**	–
JAV	**76**	95	**73**	96	**85**	–
smcfcs	**6**	97	**49**	**86**	97	–
jomo	**77**	**88**	**52**	**80**	**85**	–

Note. Biases were rounded to three decimal places. Q is the ‘true’ population parameter. OC refers to the estimates based on the available observed cases. PI is passive imputation, JAV the just another variable approach. smcfcs is substantive model compatible fully conditional specification, jomo is substantive model compatible joint modelling. β are the coefficients in the fixed part of the model, σ_u^2 denotes the intercept variance. Use of boldface type in the top part of the table indicates large absolute biases that were more than 10% the size of the true parameter. Use of boldface type in the bottom part indicates seriously low coverage, i.e. below 90%. Standard errors and confidence intervals for the random part of the model were not computed.

Discussion

This paper evaluated four different imputation strategies to fill-in incomplete predictor variables in a random intercept model. In three Monte Carlo simulations, in which the distribution of the predictors was varied (normal, heavy-tailed, skewed) the performance of these methods was compared against case deletion, which typically yields unbiased statistical inferences regarding the regression parameters, when the MCAR assumption holds—regardless of the shape of the distribution of the predictor variables.

The aim of this paper was to evaluate currently available ad hoc and state-of-the-art methods, which—unfortunately at the moment all have their strengths and weaknesses in certain situations.

The first method, passive imputation does not impute variables that are functions of other variables, but forms them passively from the imputed values of these variables. In comparison to the other methods, regardless of the distribution of the predictors, PI yielded the overall worst inferences. This result was not astonishing, since PI uses a misspecified imputation model. Present results corroborate findings from previous missing data research (e.g. Seaman et al., 2012). When adopting PI, applied researchers have to bear in mind, that bias is to be expected, when the coefficients of nonlinear terms are different from zero. Magnitude of this bias could be expected to depend on the size of this coefficient (i.e. larger effect size, larger bias).

Secondly, the JAV approach imputes variables that are functions of other variables just as any other incomplete variable in the data set. In comparison to PI, and in comparison to the more complex MI models, the simple JAV method performed quite well: In the normal predictors condition (simulation 1), parameter estimates were widely unbiased and coverage rates were usually acceptable. Only the intercept estimate was a little bit off. The most likely explanation for this finding is that the imputation model did not consider the clustered structure of the data. JAV could, however, also be applied adopting a multilevel imputation model. It needs to be noted, that the present simulation only examined an MCAR mechanism. Present findings are in line with previous research by Seaman et al. (2012), which suggests that JAV could work well under MCAR mechanisms (and only works well, when the MCAR assumption is met). Note furthermore that when predictors were not normally distributed, some bias and inacceptably low coverage rates were found. In these conditions, JAV was applied using the predictive mean matching method in mice. While pmm can be regarded as somewhat robust against violations of distributional assumptions (see the simulations in Kleinke, 2017), pmm fails, when parametric assumptions are too severely violated. This appeared to be the case in simulations 2 (heavy-tailed) and 3 (skewed incomplete predictors).

Thirdly, the conditional modelling based substantive model compatible MI approach yielded acceptable inferences (for the fixed part of the model), when data were normal, but did not perform well, when predictors were heavy-tailed or skewed. Although the intercept variance was very low (.1) in this simulation, yielding an intraclass correlation coefficient of only about 17%, ignoring these cluster effects during imputation yielded an estimate of the intercept variance that was typically too low. A huge disadvantage at the moment for applied researchers is that smcfcs is based on a single-level 'normal' imputation model and that neither the multilevel structure of the data could be considered nor that more robust imputation models could be adopted. If correct estimation of random effects is no concern, and distributional assumptions regarding the incomplete predictors are more or less met, smcfcs is a suitable imputation method.

Finally, the joint modelling software jomo produced the overall most accurate point estimates of all imputation methods. The random intercept imputation model was the only model that was fully compatible to the subsequent analysis model (at least in the first simulation). Only in the second and third simulation, when predictors were not 'normal', MI standard errors produced by jomo appeared to be too low, and coverage rates dropped below the acceptable 90% threshold, i.e. twice the nominal error rate (cf. Schafer & Graham, 2002).

Practical implications

Firstly, passive imputation cannot be recommended at all. Secondly, the JAV method can be recommended for situations, where (a) the MCAR-assumption is likely to hold, and (b) the imputation model fits the problem at hand. Note again, that JAV is not restricted to the methods that were applied in the present paper—Bayesian linear regression and predictive mean matching. Other imputation methods could be adopted that are more suitable for heavy-tailed or skewed variables (see the next section about limitations and future research for suggestions in this regard). Thirdly, smcfcs can be recommended, if distributional assumptions are met and applied researchers are mainly interested in the fixed part of the model. Finally, jomo can be recommended, and if applied researchers are mostly interested in the point estimates, jomo might also be an option, if the distribution of predictors is not entirely normal.

The next question that needs to be discussed is whether to impute at all or whether to rely on the available cases: Multiple Imputation usually is a good strategy also under MCAR mechanisms, since it prevents loss of cases and subsequently loss of statistical power. However, as the present simulation findings have shown, this recommendation can only generally be made, if—and only if all parametric modelling assumptions hold and an adequate imputation strategy is chosen. Usually the distribution of predictor variables in (multilevel) regression models is of no concern to data analysts. However, when these predictors are incompletely observed, and values are imputed by regression-based imputation techniques, the conditional distribution of these variables given other variables that shall predict missing information does become an issue.

In simulation 1, both substantive model compatible MI approaches yielded accurate fixed effects estimates as well as acceptable coverage rates. jomo also yielded an unbiased random effects variance estimate. Under MCAR, also the simple JAV stragegy yielded quite good results. However, performance of all MI methods either in bias regarding point estimates and / or regarding measures of uncertainty deteriorated, when data departed from normality.

Thus, generally speaking, when the MCAR assumption is likely to hold, predictors are not normal, and loss of power due to exclusion of cases is rather minor, then the best strategy clearly is *not* to impute. If loss of cases due to incomplete predictors is substantial and the parametric modelling assumptions regarding the imputation model are believed to hold, then imputation could be regarded as the method of choice, if, and only if the imputation method is proper. This includes that the imputation model needs to reflect the relationsships in the substantive model of scientific interest (i.e. the data analyst's model, which reflects the assumed data generating process).

Finally, one factor that also might affect the choice of the missing data strategy is computing time: for complex models and large data sets, or even 'big data' analyses, computing time will be an issue. For example, in the first simulation, passive imputation (which cannot be recomended) took only about 10 hours to run all 1000 replications on our linux server (see computational details below). The JAV method had to impute two more variables and needed about 15 hours. The substantive model compatible approaches needed considerably longer: the conditional modelling approach took a total of nearly 5 days. Finally, jomo fitted a (slightly more complex) random intercept model, and needed 6 days to run all replications.

Limitations and future research

A limitation of the present simulation is that it used a 'normal' Bayesian linear regession model in simulation 1 and 'normal' predictive mean matching in simulations 2 and 3 to apply the JAV method. The aim of the present paper was to evaluate standard ad hoc methods that can be applied in many software packages and compare them to state-of-the-art methods (that are currently available only in some packages). The mice software in R is one of the standard programms to create multiple imputations. Adaptations of the conditional modelling framework (like ice for Stata) are also found in other software packages. Usually, normal model based predictive mean matching is implemented in standard MI software. Future research could test the JAV method using imputation algorithms and models that are more robust against violations of the normal model, like non-parametric predictive mean matching variants, or semi- or non-paramatric imputation models that could be explicitly taillored to the respective non-normal situations. Two methods that could be of interest in the present scenario could be packages Qtools (Geraci, 2016) and ImputeRobust (Salfrán, 2018), which include functions to impute missing data based on quantile regression or generelalized additive models for location, scale, and shape.

Since JAV appears to work well only under MCAR, a more fruitful avenue for future research and software development will be to generalize the smcfcs approach also to multilevel imputation models, and to also look at 'non-normal' data problems. Future research also needs to identfiy, why jomo seems to produce inadquate standard error estimates, when distributional assumptions are violated, and what can be done to remedy this (e.g. by computing bootstrap standard errors). This however, will most certainly increase computing times a lot. One solution to remedy this could be to use compiled code instead of plain R code.

Note that the present simulations only looked at MCAR scenarios. Present findings should be replicated for situations where missing data are MAR. Also, the sample size was quite large. Since multilevel (imputation) models usually require a substantial sample size—which at least in the field of psychology is often hard to obtain, future research could also address the problem, how the state-of-the-art methods fare (in comparison to the far more simple ad hoc methods) when sample size is considerably smaller.

Computational details and availability of data

Results were obtained using R version 4.0.4 running under Ubuntu 18.04.5 LTS with packages mice_3.13.0, jomo_2.7-2, smcfcs_1.4.2, lme4_1.1-26 and mitml_0.4-1. Simulated data and R script files are available from https://osf.io/6d78w/.

References

Aßmann, C., Gaasch, C., Pohl, S., & Carstensen, C. H. (2015). Bayesian estimation in IRT models with missing values in background variables. Psychological Test and Assessment Modeling, 57(4), 595–618.

Barnard, J., & Rubin, D. B. (1999). Small-sample degrees of freedom with multiple imputation. Biometrika, 86(4), 948–955.

Bartlett, J. W., & Keogh, R. (2020). Smcfcs: Multiple imputation of covariates by substantive model compatible fully conditional specification. Retrieved from https://CRAN.R-project.org/package=smcfcs

Bartlett, J. W., Seaman, S. R., White, I. R., & Carpenter, J. R. (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. Statistical Methods in Medical Research, 24(4), 462–487. doi:10.1177/0962280214521348

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67(1), 1–48. doi:10.18637/jss.v067.i01

Bryk, A. S., & Raudenbush, S. W. (1992). Hierarchical linear models. Newbury Park, CA: Sage.

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

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

Committee for Medicinal Products for Human Use. (2010). Guideline on missing data in confirmatory clinical trials. London: European Medicines Agency.

Drechsler, J. (2015). Multiple imputation of missing multilevel data – rigor versus simplicity. Journal of Educational and Behavioral Statistics, 40(1), 69–95. doi:10.3102/1076998614563393

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(2), 222–240. doi:10.1037/met0000063

Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. Psychological Methods, 14(3), 275–299. doi:10.1037/a0015825

Gaffert, P., Meinfelder, F., & Bosch, V. (2018). Towards multiple-imputation-proper predictive mean matching. Proceedings of the Survey Research Methods Section of the American Statistical Association, 1026–1039.

Geraci, M. (2016). Qtools: A collection of models and tools for quantile inference. The R Journal, 8(2), 117–138.

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(2), 553–564.

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

Grund, S., Lüdtke, O., & Robitzsch, A. (2016). Multiple imputation of missing covariate values in multilevel models with random slopes: A cautionary note. Behavior Research Methods, 48(2), 640–649.

Grund, S., Lüdtke, O., & Robitzsch, A. (2018a). Multiple imputation of missing data at level 2: A comparison of fully conditional and joint modeling in multilevel designs. Journal of Educational and Behavioral Statistics, 43(3), 316–353. doi:10.3102/1076998617738087

Grund, S., Lüdtke, O., & Robitzsch, A. (2018b). Multiple imputation of missing data for multilevel models: Simulations and recommendations. Organizational Research Methods, 21(1), 111–149. doi:10.1177/1094428117703686

Grund, S., Robitzsch, A., & Luedtke, O. (2021). Mitml: Tools for multiple imputation in multilevel modeling. Retrieved from https://CRAN.R-project.org/package=mitml

Kleinke, K. (2017). Multiple imputation under violated distributional assumptions – a systematic evaluation of the assumed robustness of predictive mean matching. Journal of Educational and Behavioral Statistics, 42(4), 371–404. doi:10.3102/1076998616687084

Kleinke, K. (2018). Multiple imputation by predictive mean matching when sample size is small. Methodology, 14(1), 3–15. doi: 10.1027/1614-2241/a000141

Kleinke, K., & Reinecke, J. (2015). Multiple imputation of overdispersed multilevel count data. In U. Engel (Ed.), Techniques, data quality and sources of error (pp. 209–226). Washington, DC: Campus/The University of Chicago Press.

Kleinke, K., & Reinecke, J. (2019). countimp version 2 – A multiple imputation package for incomplete count data (Technical Report). Siegen, Germany: University of Siegen, Department of Education Studies; Psychology. Retrieved from https://kkleinke.de/countimp

Kleinke, K., Reinecke, J., Salfrán, D., & Spiess, M. (2020). Applied Multiple Imputation. Advantages, Pitfalls, New Developments and Applications in R. Cham, CH: Springer Nature.

Kleinke, K., Stemmler, M., Reinecke, J., & Lösel, F. (2011). Efficient ways to impute incomplete panel data. Advances in Statistical Analysis, 95(4), 351–373. doi:10.1007/s10182-011-0179-9

Köhler, C., Pohl, S., & Carstensen, C. H. (2015). Investigating mechanisms for missing responses in competence tests. Psychological Test and Assessment Modeling, 57(4), 472–498.

Little, R. J. A. (1988). Missing-data adjustments in large surveys. Journal of Business & Economic Statistics, 6(3), 287–296.

Lüdtke, O., Robitzsch, A., & Grund, S. (2017). Multiple imputation of missing data in multilevel designs: A comparison of different strategies. Psychological Methods, 22(1), 141–165. doi:10.1037/met0000096

Quartagno, M., & Carpenter, J. (2020). jomo: A package for multilevel joint modelling multiple imputation. Retrieved from https://CRAN.R-project.org/package=jomo

Robitzsch, A., Grund, S., & Henke, T. (2017). miceadds: Some additional multiple imputation functions, especially for mice. Retrieved from https://CRAN.R-project.org/package=miceadds

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

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. New York: Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years. Journal of the American Statistical Association, 91(434), 473–489.

Salfrán, D. & Spiess, M.. (2018). Generalized additive model multiple imputation by chained equations with package ImputeRobust. The R Journal, 10(1), 61–72.

Schafer, J. L. (1997). Analysis of incomplete multivariate data. London: Chapman & Hall.

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

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(2), 437–457. doi:10.1198/106186002760180608

Seaman, S. R., Bartlett, J. W., & White, I. R. (2012). Multiple imputation of missing covariates with non-linear effects and interactions: An evaluation of statistical methods. BMC Medical Research Methodology, 12(1), 46.

Twisk, J. W., Rijnhart, J. J., Hoekstra, T., Schuster, N. A., Ter Wee, M. M., & Heymans, M. W. (2020). Intention-to-treat analysis when only a baseline value is available. Contemporary Clinical Trials Communications, 20, 1–7. doi:https://doi.org/10.1016/j.conctc.2020.100684 van Buuren, S. (2011). Multiple imputation of multilevel data. In J. J. Hox & J. K. Roberts (Eds.), Handbook of advanced multilevel analysis (pp. 173–196). New York, NY: Taylor & Francis.

van Buuren, S. (2012). Flexible imputation of missing data. Boca Raton: CRC press.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. Journal of Statistical Software, 45(3), 1–67. doi:10.18637/jss.v045.i03

Vidotto, D., Vermunt, J. K., & Kaptein, M. C. (2015). Multiple imputation of missing categorical data using latent class models: State of the art. Psychological Test and Assessment Modeling, 57(4), 542–576.

Vink, G., Lazendic, G., & van Buuren, S. (2015). Partitioned predictive mean matching as a multilevel imputation technique. Psychological Test and Assessment Modeling, 57(4), 577–594.

von Hippel, P. T. (2009). How to impute interactions, squares, and other transformed variables. Sociological Methodology, 39(1), 265–291. doi:10.1111/j.1467-9531.2009.01215.x