

Strategies for handling missing data in longitudinal studies with questionnaires

Nazanin Noorae, Geert Molenberghs, Johan Ormel & Edwin R. Van den Heuvel

To cite this article: Nazanin Noorae, Geert Molenberghs, Johan Ormel & Edwin R. Van den Heuvel (2018) Strategies for handling missing data in longitudinal studies with questionnaires, Journal of Statistical Computation and Simulation, 88:17, 3415-3436, DOI: [10.1080/00949655.2018.1520854](https://doi.org/10.1080/00949655.2018.1520854)

To link to this article: <https://doi.org/10.1080/00949655.2018.1520854>



© 2018 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 18 Sep 2018.



Submit your article to this journal [↗](#)





Article views: 426



View Crossmark data [↗](#)

Strategies for handling missing data in longitudinal studies with questionnaires

Nazanin Noorae^a, Geert Molenberghs ^{b,c}, Johan Ormel^d and Edwin R. Van den Heuvel ^a

^aDepartment of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, Netherlands; ^bI-BioStat, Katholieke Universiteit Leuven, Leuven, Belgium; ^cI-BioStat, Universiteit Hasselt, Diepenbeek, Belgium; ^dInterdisciplinary Center of Psychopathology and Emotion Regulation, University of Groningen, University Medical Center Groningen, Groningen, Netherlands

ABSTRACT

Missing data methods, maximum likelihood estimation (MLE) and multiple imputation (MI), for longitudinal questionnaire data were investigated via simulation. Predictive mean matching (PMM) was applied at both item and scale levels, logistic regression at item level and multivariate normal imputation at scale level. We investigated a hybrid approach which is combination of MLE and MI, i.e. scales from the imputed data are eliminated if all underlying items were originally missing. Bias and mean square error (MSE) for parameter estimates were examined. ML seemed to provide occasionally the best results in terms of bias, but hardly ever on MSE. All imputation methods at the scale level and logistic regression at item level hardly ever showed the best performance. The hybrid approach is similar or better than its original MI. The PMM-hybrid approach at item level demonstrated the best MSE for most settings and in some cases also the smallest bias.

ARTICLE HISTORY

Received 29 December 2017
Accepted 4 September 2018

KEYWORDS

Fully conditional specification; latent variable models; maximum likelihood; multiple imputation

1. Introduction

A common approach in performing observational research is to collect information with multi-item questionnaires. The item scores are often used to calculate a sum score or scale in order to summarize the information or to obtain one measure for the domain of interest. However, calculation of sum scores is complicated or impossible when just one or a few items are being unanswered.

Numerous missing data methods have been developed to deal with the incomplete questionnaire data. Historically, complete case (CC) and single imputation (SI) approaches have been used frequently [1]. The CC approach uses the data on individuals with a complete set of answered items, while the SI approach substitutes the missing items or scales with a single plausible value. Although SI approaches may be relevant to complete cohort data, it has been well known that, in general, most of the SI strategies often produce biased estimates and may underestimate the standard error of the parameter estimates and hence can

CONTACT Edwin R. Van den Heuvel  e.r.v.d.heuvel@tue.nl

produce narrower confidence intervals. The limitation of CC and SI have been well discussed (see for instance [2,3]). The way to overcome these limitations is applying multiple imputation (MI).

The principle of MI is to substitute missing observations with plausible values multiple times, say M times, and create M complete data sets. Each complete data set is analysed independently and the results of the analyses are then combined using Rubin's rule [4]. Theoretically, MI is reliable when the missing mechanism fulfills the missing completely at random (MCAR) and the missing at random (MAR) assumptions [2,5]. MCAR refers to situations that the probability of missingness does not depend on the observed and missing outcomes while the MAR assumption indicates that the missingness mechanism depends on the observed outcome(s) but not on the missing outcomes [6,7]. Under these conditions, MI mostly outperforms CC and SI for questionnaire data [8–10], although the difference could be small [11].

MI can be conducted with parametric and non-parametric approaches. Parametric approaches apply either joint distributions [4,12] or conditional distributions [13,14]. The latter approach is called fully conditional specification (FCS) or multivariate imputation by chained equations (MICE) [15]. The joint distribution approach mainly applies a multivariate normal distribution (MVN) for all input variables in the imputation process irrespective of the types of variables, while FCS typically applies some kinds of regression model for each variable as a function of all other variables in the imputation process. Non-parametric imputation methods like predictive mean matching (PMM) use grouping algorithms to create donor groups from which missing data is imputed [5]. These approaches have all been applied to questionnaire data.

The MVN approach was studied on scale and item level for cross-sectional questionnaire data using expectation-maximization (EM) algorithm and Markov Chain Monte Carlo (MCMC) method [16]. The results of this study did not demonstrate a clear or significant difference between these two algorithms. Resseguier et al. [17] investigated the performance of PMM and FCS using a multinomial logit model at item level as well as the PMM imputation method at the scale level on a real data set. They recommended that either PMM or multinomial logit regression imputation at the item level should be used. Gottschall et al. [18] conducted a comparison study between item and scale imputation using regression models. They applied the EM algorithm to generate starting values for the imputation, and they did not round the imputed values to integers. Although their analyses showed negligible biases in the parameter estimates of their statistical model, the mean square error (MSE) indicated that imputation at item level increases the statistical power. Yet, another cross-sectional simulation study investigated the influence of various MI methods for a covariate that is constructed from multiple items under different missing mechanisms [19]. This study suggested that imputation at the item level with either PMM or regression model should be used before calculating the score of the covariate sum score.

An alternative approach to handling missing scales under the MAR assumption and when the scale represents an outcome measure is to just apply maximum likelihood (ML) inference. Based on the distributional properties of the statistical model, ML does not inappropriately add information from the observed data to address the missingness [2,5,20–23]. Von Hippel [24] and White et al. [3] even suggested to discard the

imputed outcomes values after the missing covariate and outcome values were substituted with MI. Their logic is that MI is needed to complete missing covariates but ML might be a better approach than MI in the presence of missing outcomes since the imputed outcomes may add noise to the parameter estimates in the final analysis [3]. It should be noted that imputation of covariates can be conducted in the same manner as imputation of outcomes. Imputation of missing covariates are essential since statistical models cannot deal with missing covariates, unless removing the corresponding case or unit.

Surprisingly, the ML approach has not been studied in details for questionnaire outcome data, although a multitude of studies compared ML with MI on non-questionnaire data [21,25–33]. Only Bell et al. [34] discussed the possibility of using ML on questionnaire data and Eekhout et al. [35] investigated the ML approach in structural equation modelling on longitudinal questionnaire data using auxiliary variables. The use of a hybrid approach, in line with [24] and [3], where the imputed scales would be discarded after MI when all items of that scale was missing, has not been studied at all. This hybrid approach would only use the imputed scales when partial information is available (through available items), but not when all items were originally missing. It can be used with all MI approaches. Therefore, the purpose of our study is to determine if there is a benefit of the hybrid approach on handling missing data in longitudinal questionnaire outcome data.

Our paper is organized as follows. The next section describes shortly the missing data approach that we will study in an extensive simulation study. The third section describes the simulation study, which is based on a real cohort case study. It describes the generation of binary items and the missing items. It also describes the statistical analysis model. This model is applied to the full data set (before imposing missingness), to the incomplete data set and to the imputed data sets for evaluation of the missing data methods. Biases and the MSE are calculated to assess the performances. The fourth section describes the simulation results and our case study, and the discussion of the results is provided in the final section.

2. Missing data methods

2.1. Multiple imputation

Since multiple imputation can be executed either on the items or on the sum scores directly, we studied both options. We considered a sum score as missing when at least one item would be missing. All imputation methods were carried out with procedure MI of the SAS software, version 9.4, under default settings [36]. To pool the obtained results of the analyses from multiple imputed data sets, we used Rubin's rule implemented in the MIANALYZE procedure in SAS [37].

Multivariate normal imputation: We applied this approach on scale level only using the MCMC approach starting with the EM algorithm to estimate the mean and covariance matrix as the starting value for the MCMC approach. The MCMC option from the MI procedure with single chain [12] and 200 burn-in iterations was employed for this method. We imposed natural boundary constraints to generate realistic scales; if the imputed sum score was outside the boundaries (i.e. below zero or above the maximum possible sum

score), we replaced the imputed sum score by the boundary value, but we did not use any further rounding to obtain integer (imputed) sums scores. This approach is referred to as the $MCMC_{scale}$ approach.

Fully conditional specification: We applied FCS imputation at the item level using logistic regression [4,14] referred to it as the LR_{item} approach in the remainder of this study. The MI procedure of SAS with the default settings uses 10 burn-in iterations. FCS for the sum scores can be performed with linear regression, but it has been noted that FCS along with linear regression and MVN provide similar results under the normality assumption [13,38]. Hence, we omitted FCS with linear regression on sum scores since we include the MVN approach on sum scores.

Predictive mean matching: We studied predictive mean matching at the item level (PMM_{item}) and at the scale level (PMM_{scale}) using the REGPMM option from the MI procedure with the default number of closest observations being 5. We implemented a constraint for imputation at the scale level: the imputed sum score using PMM was not allowed to be lower than the sum score of the available items for each individual. If PMM violated this constraint, the imputed sum score was replaced by the sum score of the available items of that particular individual.

2.2. Maximum likelihood inference

When we apply ML, we have to make a decision on when the scale would be considered as missingness. In one setting, the scales or sum scores are treated as missing when at least one item was missing but in another setting, sum scores are considered missing when all items were missing. This means that we calculated sum scores on the available items irrespective of how many items were missing, as a proportion to the number of items being available, we denoted the ML approaches ML_1 and ML_{10} , respectively.

To combine the advantages of maximum likelihood and multiple imputation, we introduced a hybrid approach at the item level and at the scale level. This hybrid approach eliminates the imputed values when all items were originally fully missing (whether imputation was conducted on item or scale level). We denoted the hybrid method by putting an 'H' in front of our notation already introduced for the imputation methods, for instance $H-PMM_{item}$ indicates the hybrid approach using PMM at item level.

3. Simulation study

The simulated incomplete data set was conducted in two steps. In the first step, a full data set questionnaire including all covariates (independent variables) as well as J binary items (dependent variables) across different follow-up times were generated. With this simulation, we tried to mimic the cohort study that motivated this research, which is described in detail in section 4.2. In the second step, binary indicators were generated at the item level to eliminate items from the full data set. The parameter settings in the simulation model were selected based on our case study and are provided in Appendix 1. We kept the parameter settings in the full data set the same in all simulations but generated different proportions of missingness and referred to as small, medium and large proportions of missingness. Each simulated data set contained 1000 individuals and setting each was repeated 500 times. Ten ($M = 10$) imputed data sets were used for MI for each repeated data set.

3.1. Full data set generation

One time-dependent covariate and five baseline covariates were generated first. The time-dependent covariate (X_{1it}) , $t = 1, 2, 3, 4$, represents age of individuals at four follow-up times. This was simulated with a multivariate normal distribution having a vector of means μ_a and a variance-covariance matrix Σ_a . One baseline binary covariate (X_{2i}) , which may indicate gender, was simulated independent of age and with a Bernoulli distribution with success probability p_g . The remaining four baseline covariates $(X_{3i}, X_{4i}, X_{5i}, X_{6i})$ were simultaneously simulated, independent of both age and gender, using a multivariate normal distribution with a vector of means μ_c and variance-covariance matrix Σ_c . Keeping only one covariate continuous, X_{6i} , three other covariates were converted into binary variables. To create these binary variables, we compared the inverse normal standard distribution function of X_{3i}, X_{4i} , and X_{5i} with pre-defined threshold probabilities p_3, p_4 , and p_5 , respectively. That is, covariate X_{qi} was converted to one if $\Phi^{-1}(x_{qi}) < p_q$, for $q = 3, 4, 5$, and Φ the cumulative standard normal distribution function. Therefore, the probability p_q , could be viewed as the success probability of binary variable X_{qi} . This way of simulating covariates guarantees that X_{3i}, X_{4i}, X_{5i} , and X_{6i} are not independent. We applied the COPULA procedure in SAS to generate these variables. The parameter settings for $\mu_a, \Sigma_a, p_g, \mu_c, \Sigma_c, p_3, p_4$, and p_5 are provided in Appendix 1.

Furthermore, we simulated 10 longitudinal correlated items ($J = 10$) using the idea of item response theory. We began by generating four correlated abilities/traits or latent variables $(Z_{i1}, Z_{i2}, Z_{i3}, Z_{i4})$ via a multivariate normal distribution having a vector of means equal to zero and a Toeplitz correlation matrix Σ_L . Then, the binary items $Y_{it}(1), Y_{it}(2), \dots, Y_{it}(J)$ for individual i at time t were drawn from a Bernoulli distribution given the latent variable Z_{it} and covariates $\mathbf{X}_{it}^T = (X_{1it}, X_{2i}, \dots, X_{6i})$. The conditional probability $\pi_{it}(j) = P(Y_{it}(j) = 1 | \mathbf{X}_{it}, Z_{it})$ is given by

$$\pi_{it}(j) = \frac{\exp(a_{tj} + b_{tj}Z_{it} + \mathbf{X}_{it}^T \mathbf{c}_{tj})}{1 + \exp(a_{tj} + b_{tj}Z_{it} + \mathbf{X}_{it}^T \mathbf{c}_{tj})}, \tag{1}$$

with \mathbf{c}_{tj} a set of coefficients for the covariates, a_{tj} and b_{tj} the difficulty and the discrimination parameters for item j at time t , respectively. The indicator Y_{it} was determined by $Y_{it}(j) = I_{[0, \pi_{it}(j)]}(U_{itj})$, with U_{itj} uniformly distributed on $(0, 1)$, and $I_A(x)$ the indicator function being equal to one when $x \in A$ and zero otherwise. The applied set of parameters a_{tj}, b_{tj} , and \mathbf{c}_{tj} in the simulation study are listed in Table A1.

3.2. Generation of missing items in the simulated full data set

We simulated a binary indicator $R_{it}(j)$ for each generated item in the full data set and eliminated $Y_{it}(j)$ from the full data set when $R_{it}(j) = 0$. We decided not to generate any missing items at the first visit, i.e. $R_{i1}(j) = 1$. Furthermore, we chose a logistic regression model for the probabilities of the binary indicator variables that depended on the covariates (\mathbf{X}_{it}) and the latent variable from the first time point (Z_{i1}) . We chose a set of parameters \tilde{a}_{tj} as the intercepts and \tilde{b}_{tj} as the slopes for the baseline latent variable, and coefficients $\tilde{\mathbf{c}}_{tj}$ for the

covariates. The probability for missing an item was set equal to

$$P(R_{it}(j) = 0) = \tilde{\pi}_{it}(j) = \frac{1}{1 + \exp\{\tilde{a}_{tj} + \tilde{b}_{tj}Z_{i1} + \mathbf{X}_{it}^T\tilde{\mathbf{c}}_{tj}\}}. \quad (2)$$

Using standard uniformly distributed random variables $U_{it}(j)$, which were independent from the uniformly distributed variables used to generate the full data set, and comparing them with the missingness probability $\tilde{\pi}_{it}(j)$ results in the missingness indicator variable, i.e. $R_{it}(j) = I_{[\tilde{\pi}_{it}(j), 1]}(U_{it}(j))$. We generated the random variables $U_{it}(j)$ in two different ways to impose either dependency or independency among the $R_{it}(j)$'s, given the latent variable and covariates. The independent setting was generated by taking independent $U_{it}(j)$ for each item j , subject i , and time points $t > 1$, whereas the dependent setting was generated by taking a uniform random variable U_{it} for all J items of subject i at time point $t > 1$. The set of parameters \tilde{a}_{tj} , \tilde{b}_{tj} , and $\tilde{\mathbf{c}}_{tj}$, used in this simulation study, are presented in Table A2.

It should be noted that the system of missing items in (2) follows an intermittent pattern of missingness rather than a drop-out pattern [39]. Furthermore, we assumed no missingness at the baseline which leads into the MAR missing mechanism with respect to the latent variables. Nevertheless, one may argue that the missing mechanism is MNAR at the level of the observed items. MNAR missingness means that the probability of being missing depends on both the unobserved and observed values. MNAR also occurs when the missingness depends on latent variable at the baseline. We explicitly generated missing items in this way since we feel that it is more realistic to have the missingness depends on the latent variable or ability of subjects at baseline instead of the response at the baseline. We strongly believe that the missingness concept is similar to answering the items themselves since a missing item represents the unanswered item with its own difficulty and discrimination parameters (\tilde{a}_{tj} and \tilde{b}_{tj}).

3.3. Statistical analysis of sum scores

To assess the impact of covariates on the repeated sum score $Y_{it} = \sum_{j=1}^J Y_{it}(j)$, it is common to select a linear-mixed model with normally distributed outcomes [40–52]. Linear-mixed models can be re-formulated as population-averaged or marginal models for normally distributed outcomes. In this case, covariates are connected to outcomes using a linear regression model and the residuals are assumed to be correlated in order to capture the associations over time.

The marginal model can be formulated as

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i, \quad (3)$$

with $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \dots, Y_{iT})$ as the vector of sum scores for individual i across time, $\mathbf{X}_i = (\mathbf{X}_{i1}, \mathbf{X}_{i2}, \dots, \mathbf{X}_{iT})$ the matrix of covariates, $\boldsymbol{\beta}$ the set of unknown regression parameters, and $\boldsymbol{\epsilon}_i$ a vector of residuals having a normal distribution with mean 0 and variance-covariance matrix \mathbf{V} , i.e. $\boldsymbol{\epsilon}_i \sim N(0, \mathbf{V})$. Various structures can be considered for the matrix \mathbf{V} , however, we choose the Toeplitz covariance matrix since this choice seems to fit appropriately to our case study. Procedure MIXED of the SAS software, version 9.4, was used to implement this model. Note that parameters in model (3) can be estimated with maximum

likelihood and restricted maximum likelihood estimation methods, though maximum likelihood is used for the current study.

All the methods for handling missing data were compared with an analysis of the full data set. We applied the same statistical model (3) on the full, imputed and incomplete data and investigated the parameter estimates in terms of bias and (square root of the) MSE. The bias indicates the mean distance of the parameter estimates from the missing data methods and the parameter estimates of the full data sets when averaged over all our simulations while MSE indicates how close the estimates are across simulations. Furthermore, we computed an average relative efficiency [4] for each parameter across imputed data sets to evaluate the imputation process with 10 imputed data sets.

4. Results

4.1. Simulation study

Table 1 provides the percentages of missing items (item missing) and the percentages of subjects who did not record any items (unit missingness) for each follow-up time (visit) for the different setting. The proportion of unit missingness is substantial for the dependent missingness indicator variables but almost non-existent for the independent missing indicator variables. The percentage of missing items is really large for the dependent missingness indicator variables.

As it was expected, imputation at the item level takes much more time than MI at the scale level such that MI at the item level could take up to 36 h, while MI at the scale level lasts almost 2 h to create 10 imputed data sets for all 500 simulated data sets. PMM at the item level is slightly faster than logistic regression, but it is slower than MCMC when it was applied at the scale level. In our simulation, generating 10 imputed data sets indicated that the average relative efficiency [4] for all parameters per simulation scenario was at least 95%, except for a few cases. The smallest average relative efficiency of 91% was observed for the correlation coefficients and variance components. It was observed with MI at the item level using logistic regression (LR_{item}) with a medium proportion of missingness for the dependent missingness indicator variables. We also observed a minimum of 92% efficiency in the regression parameters for LR_{item} when the proportion of missingness was large.

Investigating the bias in the fixed effects showed that the logistic regression on items and the two imputation methods on scales, with or without the hybrid approach, hardly ever obtained the smallest bias compared with the other four methods. Among these four missing data methods, there is no clear winner for the fixed effect estimates, since they fluctuate in their rankings on the smallest bias across parameters and missing data settings. For instance, the smallest absolute bias for the fixed effect of the second follow-up time (β_{T_2})

Table 1. The overall proportion (percentage) of missingness in the simulated data sets.

Missingness indicator variable	Proportion of missingness	Unit missing			Item missing		
		Visit 2	Visit 3	Visit 4	Visit 2	Visit 3	Visit 4
Independent	Small	0.004	0.72	1.46	4.19	13.95	10.15
	Medium	0.000	1.75	8.80	7.20	26.23	24.04
Dependent	Medium	4.95	23.43	23.56	7.26	26.33	24.06
	Large	8.93	45.73	37.06	12.75	53.12	44.03

is obtained by ML_{10} , $H-PMM_{item}$, PMM_{item} , and ML_1 , respectively, when a large number of missing items was generated (Table 2). However, this order is fully reversed for the same parameter when only a small number of items are missing. On the other hand, it seems that the maximum likelihood methods provide somewhat smaller biases, in general, for the fixed effects parameters that are related to follow-up times. For the bias in correlation coefficients, the maximum likelihood method ML_{10} performs best for large numbers of missing items, while $H-PMM_{item}$ performs best for all other settings of generating missing data.

Testing for no biases with the Wilcoxon signed rank test demonstrates that bias in almost all parameters for all methods and under all settings are significant at the level of $\alpha = 0.05$. The bias with respect to the estimate of the full data set is not just significant but can also be substantial even for the best performing methods. For instance, the absolute bias for parameter β_{T_3} is 0.792 with ML_{10} (Table 2) when large proportions of missing items were generated. The parameter β_{T_3} is estimated at -2.375 with the full data set, which implies a bias of 33% bias with ML_{10} . In other words, the missing data methods are not fully unbiased when substantial proportions of items are missing.

The simulation results on MSE demonstrate that the hybrid method $H-PMM_{item}$ performs generally best on almost all fixed effects parameters and on the correlation parameters in all our settings compared with all other methods. For instance, the square root MSE is determined at 0.783 for an average estimate of 1.411 in the full data set for the interaction effect of gender and follow-up time (β_{2T_4}) in the setting with large numbers of missing items (Table 3). The other methods are closer to or even above a square root MSE of one. When $H-PMM_{item}$ is outperformed by another method for a specific parameter, the hybrid method is still close to the other method (relative to the estimate in the full data set). Even stronger, we could see that the hybrid methods $H-PMM_{item}$ and $H-LR_{item}$ outperform their original methods PMM_{item} and LR_{item} , respectively, on almost all fixed effects and correlation coefficients, although differences between the hybrid and the original method are never very large relative to the estimates in the full data set. Contrary to this, the hybrid approach at the scale level does not seem to contribute or is even somewhat worse than the corresponding MI method. MSE does not reveal a clear winner for the estimation of variance components. The worst performing methods on variance components are in general the hybrid MCMC method on scales ($H-MCMC_{scale}$) and the maximum likelihood approach (Tables 3 and 4).

4.2. The TRAILS cohort: a motivating example

The Dutch cohort ‘Tracking Adolescents’ Individual Lives Survey: TRAILS’ is a longitudinal study to assess the development of mental health from childhood towards adulthood [53]. Data on the psychological, social and physical health of 2230 participants (49% boys and 51% girls) at age 10–12 was collected through a special questionnaire, and they were asked to participate in bi-annual follow-ups until they reached an age of 24. In the current study, we decided to study the depression domain of the first four waves. This domain is investigated with a questionnaire containing 13 items for the first three waves (youth self-report) and 18 items for the fourth wave (adolescent self-report). This change in questionnaire is due to the psychological changes during adolescence. All relevant items were designed as multiple choice with three levels (0, 1, 2). We only used the 10 common items in all four waves and treated the items as binary responses (zero and non-zero).

Table 2. Absolute bias of the fixed effect estimates, with large proportion of missingness.

	Parameter	Full data set	ML_{10}	ML_1	PMM_{item}	LR_{item}	PMM_{scale}	$MCMC_{scale}$	$H-PMM_{item}$	$H-LR_{item}$	$H-PMM_{scale}$	$H-MCMC_{scale}$
Fixed effects	β_0	25.413	-1.382*	-1.20*	-0.47*	-2.032*	-1.079*	-0.902*	-0.928*	-1.220*	-1.152*	-1.255*
	β_1	0.007	0.128*	0.12*	0.04*	0.192*	0.095*	0.082*	0.089*	0.117*	0.109*	0.117*
	β_2	-5.592	0.002	0.00	0.00*	0.001*	0.000*	0.000*	0.000*	0.000*	0.000*	0.001*
	β_3	0.736	-0.061*	-0.12*	0.08*	-0.048*	0.067*	0.004*	-0.062*	-0.088*	-0.071*	-0.047*
	β_4	-0.673	0.079*	0.14*	-0.23*	-0.207*	-0.088*	-0.008*	0.034*	0.037*	0.071*	0.060*
	β_5	1.449	-0.173*	-0.32*	0.04*	-0.435*	-0.057*	-0.114*	-0.183*	-0.266*	-0.237*	-0.269*
	β_6	0.293	0.124*	0.01	0.13*	0.271*	0.227*	0.142*	0.024*	0.025*	0.069*	0.222*
Fixed effects of time	β_{T2}	0.026	0.004	-0.95*	-0.54*	-0.791*	-1.220*	-1.175*	-0.521*	-0.603*	-1.084*	-1.304*
	β_{T3}	-2.375	0.792*	-1.67*	-1.69*	-3.198*	-2.153*	-2.097*	-1.453*	-1.785*	-1.975*	-2.678*
	β_{T4}	-5.135	1.789*	-1.12*	-1.20*	-3.107*	-1.515*	-1.507*	-1.006*	-1.357*	-1.377*	-1.734*
Interaction of gender with time	β_{2T_2}	-3.024	0.044*	0.35*	-0.18*	0.146*	0.117*	0.124*	0.103*	0.195*	0.212*	0.285*
	β_{2T_3}	-1.081	-0.744*	0.25*	-1.54	-0.089*	-1.196*	-0.313*	-0.096*	0.179*	-0.077*	-0.873
	β_{2T_4}	1.411	-0.669*	-0.16*	-0.98*	-0.514*	-0.891*	-0.487*	-0.308*	-0.222*	-0.331*	-0.725*
Correlation	ρ_{12}	0.399	0.021	0.041	0.025	0.029	0.045	0.044	0.024	0.024	0.049	0.040
	ρ_{13}	0.310	0.059	0.086	0.099	0.146	0.099	0.097	0.076	0.090	0.124	0.094
	ρ_{14}	0.241	0.073	0.061	0.084	0.085	0.077	0.068	0.059	0.062	0.079	0.070
	ρ_{23}	0.383	-0.021	0.068	0.078	0.166	0.084	0.082	0.062	0.081	0.096	0.075
	ρ_{24}	0.313	0.023	0.049	0.064	0.099	0.066	0.055	0.048	0.057	0.064	0.057
	ρ_{34}	0.376	0.037	0.036	0.078	0.049	0.071	0.047	0.044	0.041	0.054	0.050
Variance components	σ_1^2	478.49	0.360	0.480	-0.40	-0.633	-0.738	-0.837	0.216	0.223	0.008	0.06
	σ_2^2	366.79	-22.670	-13.630	5.85	-2.076	2.352	4.304	-4.491	-7.022	-3.183	-6.39
	σ_3^2	239.20	-108.050	-49.390	19.64	-44.616	-3.404	-16.494	-23.326	-39.079	-26.409	-31.77
	σ_4^2	220.38	-69.910	-30.690	13.22	-61.465	-4.072	-5.334	-14.216	-30.728	-18.071	-20.43

* Significate bias in the level of $\alpha = 0.05$ using Wilcoxon signed rank test.

**Table 3.** Square root of MSE for the large proportion of missingness.

	Parameter	Full data set	ML_{10}	ML_1	PMM_{item}	LR_{item}	PMM_{scale}	$MCMC_{scale}$	$H-PMM_{item}$	$H-LR_{item}$	$H-PMM_{scale}$	$H-MCMC_{scale}$
Fixed effects	β_0	25.413	5.308	5.413	4.420	6.779	5.943	6.378	4.322	4.658	5.303	5.052
	β_1	0.007	0.479	0.491	0.399	0.614	0.536	0.576	0.391	0.422	0.480	0.457
	β_2	-5.592	0.032	0.034	0.030	0.044	0.038	0.042	0.027	0.029	0.033	0.033
	β_3	0.736	0.449	0.472	0.412	0.579	0.507	0.563	0.371	0.400	0.458	0.438
	β_4	-0.673	0.792	0.793	0.722	1.033	0.837	0.934	0.651	0.698	0.779	0.740
	β_5	1.449	0.658	0.714	0.532	0.928	0.688	0.768	0.548	0.607	0.668	0.654
	β_6	0.293	0.297	0.300	0.269	0.446	0.391	0.398	0.232	0.245	0.294	0.345
Fixed effects of time	β_{T2}	0.026	0.296	1.018	0.594	0.857	1.266	1.230	0.580	0.657	0.992	1.345
	β_{T3}	-2.375	1.209	1.929	1.795	3.480	2.319	2.306	1.626	1.970	2.041	2.825
	β_{T4}	-5.135	1.901	1.320	1.299	3.457	1.652	1.665	1.127	1.549	1.410	1.852
Interaction of gender and time	β_{2T_2}	-3.024	0.372	0.559	0.374	0.398	0.441	0.468	0.332	0.381	0.473	0.502
	β_{2T_3}	-1.081	1.319	1.197	1.709	1.146	1.594	1.169	0.879	0.969	1.176	1.383
	β_{2T_4}	1.411	1.052	0.949	1.178	1.344	1.232	1.013	0.783	0.823	1.001	1.116
Correlation	ρ_{12}	0.399	0.023	0.042	0.026	0.031	0.046	0.046	0.025	0.025	0.050	0.045
	ρ_{13}	0.310	0.067	0.093	0.102	0.154	0.105	0.104	0.081	0.094	0.128	0.102
	ρ_{14}	0.241	0.035	0.075	0.081	0.177	0.090	0.088	0.066	0.087	0.100	0.084
	ρ_{23}	0.383	0.078	0.068	0.087	0.092	0.083	0.075	0.063	0.067	0.084	0.075
	ρ_{24}	0.313	0.032	0.054	0.067	0.108	0.070	0.067	0.052	0.061	0.068	0.060
	ρ_{34}	0.376	0.050	0.050	0.082	0.068	0.078	0.060	0.051	0.050	0.063	0.057
Variance components	σ_1^2	478.49	0.693	0.796	0.711	1.051	1.108	1.223	0.547	0.579	0.644	0.660
	σ_2^2	366.79	23.623	15.001	7.188	5.891	6.058	7.049	6.161	8.276	6.347	6.647
	σ_3^2	239.20	109.861	52.753	22.276	52.284	15.296	21.380	26.605	41.456	30.600	38.356
	σ_4^2	220.38	71.919	33.530	15.888	70.411	12.088	11.645	17.479	33.357	21.437	23.235

Table 4. Square root of MSE for the medium proportion of missingness.

	Parameter	Full data set	ML_{10}	ML_1	PMM_{item}	LR_{item}	PMM_{scale}	$MCMC_{scale}$	$H-PMM_{item}$	$H-LR_{item}$	$H-PMM_{scale}$	$H-MCMC_{scale}$
Fixed effects	β_0	25.413	6.025	3.794	2.445	2.784	5.465	7.052	2.392	2.560	5.466	7.360
	β_1	0.007	0.545	0.342	0.220	0.252	0.495	0.637	0.142	0.185	4.491	1.506
	β_2	-5.592	0.037	0.023	0.016	0.019	0.033	0.042	0.015	0.017	0.033	0.044
	β_3	0.736	0.507	0.334	0.219	0.239	0.457	0.608	0.209	0.223	0.454	0.633
	β_4	-0.673	0.886	0.577	0.381	0.414	0.793	1.094	0.358	0.373	0.785	1.147
	β_5	1.449	0.315	0.202	0.167	0.223	0.384	0.429	0.287	0.323	0.747	0.938
	β_6	0.293	0.778	0.446	0.295	0.365	0.725	0.886	0.155	0.185	0.390	0.429
Fixed effects of time	β_{T2}	0.026	1.478	0.158	0.142	0.185	4.491	1.906	0.464	0.672	4.802	2.461
	β_{T3}	-2.375	2.335	0.530	0.473	0.689	4.771	2.813	0.497	0.822	2.503	1.302
	β_{T4}	-5.135	1.296	0.454	0.542	1.089	2.420	1.592	0.216	0.232	0.495	0.666
Interaction of gender and time	β_{2T_2}	-3.024	1.153	0.206	0.198	0.185	1.189	1.006	0.198	0.185	1.189	1.214
	β_{2T_3}	-1.081	1.590	0.663	0.585	0.421	1.193	1.471	0.551	0.411	1.187	1.701
	β_{2T_4}	1.411	0.834	0.697	0.584	0.646	0.899	0.905	0.518	0.539	0.906	0.888
Correlation	ρ_{12}	0.399	0.008	0.047	0.005	0.005	0.059	0.050	0.005	0.005	0.059	0.048
	ρ_{13}	0.310	0.066	0.095	0.028	0.036	0.119	0.107	0.027	0.034	0.012	0.101
	ρ_{14}	0.241	0.047	0.073	0.037	0.076	0.113	0.084	0.034	0.056	0.119	0.073
	ρ_{23}	0.383	0.076	0.078	0.024	0.025	0.098	0.088	0.023	0.025	0.099	0.085
	ρ_{24}	0.313	0.061	0.057	0.030	0.052	0.079	0.064	0.028	0.039	0.080	0.059
	ρ_{34}	0.376	0.097	0.055	0.034	0.043	0.081	0.067	0.032	0.037	0.081	0.063
Variance components	σ_1^2	478.49	0.828	0.475	0.336	0.370	1.075	1.497	0.297	0.310	1.038	1.485
	σ_2^2	366.79	30.673	11.881	5.321	3.684	11.194	19.364	5.364	3.681	11.195	35.417
	σ_3^2	239.20	59.341	97.360	12.362	8.692	17.245	26.322	11.173	8.904	17.149	65.150
	σ_4^2	220.38	28.553	92.536	9.581	15.097	8.814	10.178	5.408	12.548	8.670	30.872

Note: The third column represents the parameter estimates for the full data set.

Table 5. Pattern of missing during the follow-up of the TRAILS

Pattern	Wave 1	Wave 2	Wave 3	Wave 4	%Complete set of items	%Available items
1	X	X	X	X	61.06	65.91
2	X	X	.	.	11.11	11.48
3	X	X	.	X	9.41	8.54
4	X	X	X	.	7.31	7.55
5	X	.	.	.	4.07	3.98
6	.	X	X	X	3.02	0.36
7	X	.	X	X	1.42	0.63
8	X	.	.	X	1.05	0.86
9	.	X	.	X	0.82	–
10	.	X	.	.	0.73	0.32

Patterns of missingness for the sum scores are provided in Table 5. In this table, the signs ‘X’ and ‘.’ indicate observed and missing outcomes, respectively. Moreover, the percentage of participants are provided based on two approaches: one column indicates the percentage of participants with a complete set of items at each wave, and the other column provides the percentage of participants that have at least one item (out of 10 items) available.

For instance, 61.06% of participants answered all 10 items at all four waves, while 11.11% of participants responded all 10 items present at the first two waves but missed at least one item at the other two waves. Furthermore, 7.55% of all participants answered at least one item available at the first three waves, but none of the items available at the fourth wave.

To study the risk factors for depression status and development, data were collected on the history of parental internalizing and externalizing behaviour, family structure, and social-economic status of the family, next to age and gender of the participants. To understand the association of these factors with depression, we studied their relation with the depression scale, i.e. the mean score of 10 items expressed as percentages. We applied the two maximum likelihood methods and the PMM at the item level with and without the hybrid approach, since these four methods were most promising in the simulation study.

Table 6. Parameter estimates (standard errors) of the real data set.

	Parameter	ML_{10}	ML_1	PMM_{item}	$H-PMM_{item}$
Fixed effects	Intercept	39.229*(8.398)	36.771*(8.620)	37.538*(8.383)	38.187*(8.334)
	Gender	–5.740*(1.146)	–5.478*(1.126)	–5.555*(1.133)	–5.497*(1.125)
	Age at baseline	–0.759 (0.755)	–0.554 (0.776)	–0.614 (0.756)	–0.672 (0.749)
	Externalising behaviour	–0.400 (1.105)	–0.837 (1.133)	–0.339 (1.092)	–0.351 (1.090)
	Internalising behaviour	2.075*(0.544)	2.240*(0.560)	2.030*(0.521)	2.045*(0.539)
	Social-economic status	–0.104 (0.546)	–0.533 (0.560)	–0.299 (0.535)	–0.159 (0.540)
	Family structure	1.950 (1.094)	2.157 (1.124)	2.026 (1.094)	1.995 (1.082)
Fixed effects of time	Time ₂	3.601*(0.818)	4.445*(0.816)	3.706*(0.811)	3.713*(0.807)
	Time ₃	1.254 (0.935)	1.453 (0.920)	0.792 (0.943)	1.350 (0.921)
	Time ₄	–2.008 (0.916)	0.084 (1.019)	–1.760 (0.908)	–1.953*(0.909)
Interaction of gender and time	Gender*Time ₂	–7.136*(1.179)	–7.527*(1.173)	–7.304*(1.146)	–7.357*(1.155)
	Gender*Time ₃	–9.790*(1.359)	–9.911*(1.335)	–9.006*(1.247)	–9.931*(1.335)
	Gender*Time ₄	–5.356*(1.341)	–3.054*(1.489)	–4.834*(1.316)	–5.401*(1.330)
Variance components	σ_1^2	675.13*(21.141)	674.43*(20.680)	671.106*(20.703)	672.612*(20.664)
	σ_2^2	620.60*(19.587)	667.63*(20.741)	615.468*(19.198)	619.058*(19.428)
	σ_3^2	585.14*(20.535)	588.40*(20.499)	564.072*(19.633)	585.522*(20.316)
	σ_4^2	531.54*(18.279)	753.31*(25.542)	513.644*(16.838)	526.272*(18.187)

* Significantly different from zero in the level of $\alpha = 0.05$.

Results of these analyses are presented in Table 6. Comparing the four methods of handling missing data demonstrates that the estimates are not identical. For most of the parameters, the differences are relatively small compared to the size of the estimate, although some substantial differences also occur. For instance, PMM_{item} gives a different effect for the third time point with respect to the other three methods. For the effect of the fourth time point and the effect of socio-economic status, ML_1 deviates strongly from the other three methods. Moreover, it seems that the maximum likelihood ML_1 to deviate for most parameters from the other three. In our example, there was only one parameter where the methods conflict in significance. The hybrid PMM method at the item level demonstrates a significant effect of time point four, while the other methods do not. Considering our simulation study, this hybrid method would most likely be the closest to the value of a complete data set.

5. Discussion

The aim of the present study was to determine how well ML would handle missing outcomes from longitudinal questionnaire data in comparison with ML at the item level or at the scale level. Multiple imputation at the item level was conducted using logistic regression via the FCS approach, and using PMM; whereas imputation at scale level was achieved using PMM and the multivariate normal imputation (MVN) using the MCMC approach. In addition to the aforementioned missing data methods, we were particularly interested in the advantage of combining MI and ML, referred to as the hybrid approach. This hybrid approach first imputes all the unobserved values (depending on the choice of imputation at the item level or at the scale level) and then eliminates the imputed scales whenever all items corresponding to those scales were originally missing. An extensive simulation study was conducted, imposing intermittent missingness with different proportions. The scale in each data set (i.e. full, incomplete, and imputed data sets) was analysed with a population-averaged model (or marginal model) and the accuracy of parameter estimates from each missing data method was compared to the parameter estimates of the full data set, in terms of bias and MSE. The relevance of our comparisons was motivated by a Dutch cohort data set (TRAILS).

Results of our simulation studies showed that MI at item level outperforms imputation at the scale level. This finding is consistent with cross-sectional studies [17,18], although it would have been possible that the information from other time points would enhance MI at scale level more than at the item level. In fact, the reverse seems true. Due to ignoring (partial) items in scale-level methods, we lost a lot of information on scale level that could not be compensated by information of scales from other time points. Nevertheless, the available information on scales from other time points was sufficient for ML since it was the best approach for estimation of the correlation coefficients across time in several settings. A study on missing data with structural equation modeling also indicated the advantage of using score information from other time points as auxiliary variables in maximum likelihood inference [35].

Another result of the current study is that the hybrid approach did not show massive differences with their original MI method with respect to the size of the estimates in the full data set. However, the hybrid approach was somewhat better when the imputation procedure was executed at item level. More importantly, the hybrid approach with PMM

at item level revealed smaller MSE compared to the other applied techniques for almost all parameters and in most of the simulation settings. This means that the parameter estimates of this approach are in general closest to the parameter estimates of the full data compared to other missing data methods. For cases where it was not the best approach, it was quite comparable to the best approach. This finding is in accordance with Von Hippel [24] and White et al. [3] who advocated to ignore the imputed dependent variables before performing the final analysis, because they add noise to the estimators. Apparently, this argument can be extended to incomplete questionnaire data. Indeed, the hybrid method overcomes the limitation of the maximum likelihood inference, which cannot handle partial available items. Imputing partially missing items improves the analysis, but fully missing sum scores do not have to be imputed since ML handles this appropriately. Hence, the partial loss of information is corrected with MI but imputing unit missingness adds noise to the parameter estimates of the analysis.

The scope of this study was limited to methods that can handle general missingness patterns and which have been established in the past. We used MVN at the scale level since it was assumed that mean scores are normally distributed across time (at least in the analysis of the scales). However, we avoided MVN at the item level since prior studies on non-questionnaire data have suggested not to use MVN when the input variables strongly violate normality assumptions [13,54–58]. In addition, Horton et al. [59] have demonstrated bias in the parameter estimates when the imputed values of the MVN are rounded to binary variables. Therefore, we imputed unobserved items using logistic regression. However, it might not result in the best method (with or without the hybrid approach) since logistic regression does not consider a subject-specific latent (ability) variable in the model. Hence, logistic regression did not fully match with the simulated item response theory model since MI does not consider a latent variable for the items [20, Chapter 2]. To overcome this limitation, we applied PMM since it retains the distribution of variables, is robust to transformation, and is less sensitive to mis-specification of the model [14]. Hence, we explored it at the item and at the scale level. Our simulation study confirmed these published advantages, although we showed that PMM is best used with the hybrid approach for questionnaire data.

The strengths of this study is the realistic longitudinal simulation model. Previous studies utilized either real data sets or Monte Carlo simulations for cross-sectional data, and a multivariate normal distribution for items in structural equation models. We simulated longitudinal questionnaire data borrowing ideas from item response theory and also took into account effects of covariates. We generated correlated latent variables across time, and specified different item characteristics (difficulty and discrimination) for each binary item over time. We selected the parameter settings based on our real motivating case study, but we did not examine the impact of possible other realistic settings for our simulation. It should be noted that our simulation model mimics longitudinal questionnaire data from just one domain, but it can also be viewed as multi-domain cross-sectional questionnaire data, where the trait of each domain is mimicked by our multivariate latent variable over time. We studied missing data methods that treat all items simultaneously rather than implementing missing data methods per time point separately. This choice is supported by Graham et al. [60] and Gottschall et al. [18] who advocated not to split domains in handling missing data, although Graham et al. [61] suggested earlier to treat the domains separately.

A weakness of our study is the dependence of the missingness indicator variables on the baseline latent variable. This introduces an MNAR missing mechanism. However, there are a few arguments for our choice. First and foremost, we are convinced that the trait of a subject may affect the missingness instead of the baseline observed score. The ability or latent variable is the true subject characteristic, while the sum score represents a somewhat arbitrary value since it could be different when observed almost immediately again. Hence, a missing item represents the inability to answer the question. Secondly, MAR can never be demonstrated and it is probably seldom fully satisfied in practice. Although our simulation produces MNAR, it is MAR at the latent level. This means that our MNAR is not that far away from MAR and may actually be a realistic MNAR model. Finally, some research is currently implementing MAR missing data methods without really knowing if MAR is satisfied. Understanding the performance of these methods in settings that are MNAR, but not dramatically violating MAR, seems reasonable since this may mimic real studies even better.

A potential topic for further research is maximum likelihood in combination with weighted approaches. Indeed, if scales could be calculated on the basis of the observed items, the precision would be determined by the number of available items. Taking into account these differences in weights, we may actually improve the maximum likelihood approach. Note that some theoretical work has been conducted on weighted maximum likelihood methods [62–65]. Alternatively, inverse probability weighting methods may also be investigated in this setting [66–69], but it is probably somewhat complicated for missing data patterns other than drop-out. Furthermore, it would be of interest to investigate if a sequential approach of imputation would improve our investigated methods. One could start imputing the first longitudinal observations using only the covariates and then sequentially impute the outcomes at the next time points using the covariates and the previous imputed outcomes. This approach would fit better with the way that the data were collected, but whether this approach would really be better is not immediately clear, since the correlation between an outcome at later time points and an outcome at an earlier time point can be informative for the imputation method when an outcome at an earlier time point is imputed. Finally, developing multiple imputation methods for questionnaire data that uses latent variable models may improve the current method and may be more realistic with respect to this type of data.

In summary, handling missing data in longitudinal questionnaire type of outcome data seems to perform best with a PMM approach at item level when the imputed unit missing scales are eliminated (the hybrid approach). This method outperformed other approaches, in particular on the fixed effects parameters of the marginal linear-mixed model. For estimation of variance components, no clear winner among missing data methods was observed. The multiple imputation methods at the scale level performed worse across almost all parameters.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Geert Molenberghs  <http://orcid.org/0000-0002-6453-5448>

Edwin R. Van den Heuvel  <http://orcid.org/0000-0001-9157-7224>

References

- [1] Eekhout I, de Boer RM, Twisk HC J W de Vet, et al. Missing data: a systematic review of how they are reported and handled. *Epidemiology*. 2012;23:729–732.
- [2] Molenberghs G, Kenward M. *Missing data in clinical studies*. Chichester: Wiley; 2007.
- [3] White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med*. 2011;30:377–399.
- [4] Rubin DB. *Multiple imputation for nonresponse in survey*. New Jersey: Wiley; 1987.
- [5] Carpenter JR, Kenward MG. *Multiple imputation and its application*. Chichester: Wiley; 2013.
- [6] Rubin DB. Inference and missing data. *Biometrika*. 1976;63:581–592.
- [7] Council NR. *The prevention and treatment of missing data in clinical trials*. Washington, DC: The National Academies Press; 2010.
- [8] Gmel G. Imputation of missing values in the case of a multiple item instrument measuring alcohol consumption. *Stat Med*. 2001;20(15):2369–2381.
- [9] van Buuren S. Item imputation without specifying scale structure. *Methodol Eur J Res Methods Behav Social Sci*. 2010;6:31–36.
- [10] Parent MC. Handling item-level missing data: simpler is just as good. *Couns Psychol*. 2013;41(4):568–600.
- [11] Shrive FM, Stuart H, Quan H, et al. Dealing with missing data in multi-question depression scale: a comparison of imputation methods. *BMC Med Res Methodol*. 2006;6:57.
- [12] Schafer JL. *Analysis of incomplete multivariate data*. London: Chapman & Hall; 1997.
- [13] van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Method Med Res*. 2007;16:219–242.
- [14] van Buuren S. *Flexible imputation of missing data*. Chapman & Hall/CRC Interdisciplinary Statistics; 2010.
- [15] van Buuren S, Groothuis-Oudshoorn S. mice: Multivariate imputation by chained equations in r. *J Stat Softw*. 2011;45:1–67.
- [16] Lin TH. A comparison of multiple imputation with EM algorithm and MCMC method for quality of life missing data. *Qual Quant*. 2010;44(2):277–287.
- [17] Resseguier N, Verdoux H, Giorgi H, et al. Dealing with missing data in the center for epidemiologic studies depression self-report scale: a study based on the French E3N cohort. *BMC Med Res Methodol*. 2013;13(28):28.
- [18] Gottschall AC, West SG, Enders CK. A comparison of item-level and scale-level multiple imputation for questionnaire batteries. *Multivariate Behav Res*. 2012;47(1):1–25.
- [19] Eekhout I, de Vet HCW, Twisk JWR, et al. Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. *J Clin Epidemiol*. 2014;67(3):335–342.
- [20] Fitzmaurice G, Davidian M, Verbeke G. *Longitudinal data analysis*. Boca Raton: CRC Press; 2009.
- [21] Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedure. *Psychol Methods*. 2001;6(4):330–351.
- [22] Jansen I, Beunckens C, Molenberghs G, et al. Analyzing incomplete discrete longitudinal clinical trial data. *Stat Sci*. 2006;21(1):52–69.
- [23] Enders CK. The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychol Methods*. 2001;6(4):352–370.
- [24] Von Hippel PT. Regression with missing y's: an improved strategy for analyzing multiply imputed data. *Sociol Methodol*. 2007;37(1):83–117.
- [25] Newman AD. Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood and multiple imputation. *Organ Res Methods*. 2003;6:328.
- [26] Schafer JL. Multiple imputation in multivariate problems when the imputation and analysis models differ. *Stat Neerl*. 2003;57(1):19–35.
- [27] Graham JW, Olchowski AE, Gilreath TD. How many imputations are really needed? some practical clarifications of multiple imputation theory. *Prev Sci*. 2007;8:206–213.

- [28] Olinsky A, Chen S, Harlow L. The comparative efficacy of imputation methods for missing data in structural equation modeling. *Eur J Oper Res.* 2003;151(1):53–79.
- [29] Ibrahim JG, Chen MH, Lipsitz SR, et al. Missing-data methods for generalized linear models. *J Am Stat Assoc.* 2005;100(469):332–346.
- [30] Messer K, Natarajan L. Maximum likelihood, multiple imputation and regression calibration for measurement error adjustment. *Stat Med.* 2008;27(30):6332–6350.
- [31] Schlomer GL, Bauman S, Card NA. Best practices for missing data management in consulting psychology. *J Couns Psychol.* 2010;57:1–10.
- [32] Larsen R. Missing data imputation versus full information maximum likelihood with second-level dependencies. *Struct Equ Model.* 2011;18:649–662.
- [33] Yuan KH, Yang-Wallentin F, Bentler PM. MI versus mi for missing data with violation of distribution conditions. *Social Methods Res.* 2012;41(4):598–629.
- [34] Bell M, Fairclough D. Practical and statistical issues in missing data for longitudinal patient reported outcomes. *Stat Method Med Res.* 2014;23(5):440–459.
- [35] Eekhout I, Enders CK, Twisk JWR, et al. Analyzing incomplete item scores in longitudinal data by including item score information as auxiliary variables. *Struct Equ Model.* 2015;50(5):504–519.
- [36] SAS Institute I. *Sas/stat 9.3 user's guide: the MI procedure.* Cary (NC): SAS Institute Inc.; 2011.
- [37] SAS Institute I. *Sas/stat 9.3 user's guide: The mianalyze procedure.* Cary (NC): SAS Institute Inc.; 2011.
- [38] Lee KJ, Carlin JB. Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. *Am J Epidemiol.* 2010;171:624–632.
- [39] Diggle P, Kenward MG. Informative drop-out in longitudinal data analysis. *J Roy Stat Soc Ser C (Appl Stat).* 1994;43:49–93.
- [40] Vinkers DJ, Gussekloo J, Stek ML, et al. Temporal relation between depression and cognitive impairment in old age: prospective population based study. *Br Med J.* 2004;329(7471):881–883.
- [41] Zarate Jr, Quiroz JA, Singh JB, et al. An open-label trial of the glutamate-modulating agent riluzole in combination with lithium for the treatment of bipolar depression. *Biol Psychiatry.* 2005;57(4):430–432.
- [42] Netemeyer RG, Maxham JG, Pulling C. Conflicts in the work-family interface: links to job stress, customer service employee performance, and customer purchase intent. *J Mark.* 2005;96(2):130–143.
- [43] Forbes-Thompson S, Gajewski B, Scott-Cawiezell J, et al. An exploration of nursing home organizational processes. *West J Nurs Res.* 2006;28(8):935–954.
- [44] Sanacora G, Kendell SF, Levin Y, et al. Preliminary evidence of riluzole efficacy in antidepressant-treated patients with residual depressive symptoms. *Biol Psychiatry.* 2007;61(6):822–825.
- [45] Slort W, Blankenstein AH, Schweitzer BPM, et al. Effectiveness of the palliative care “availability, current issues and anticipation (aca) communication” training programme for general practitioners on patient outcomes: a controlled trial. *Palliat Med.* 2014;28(9):1036–1045.
- [46] Walsemann KM, Bell BA, Maitra D. The intersection of school racial composition and student race/ethnicity on adolescent depressive and somatic symptoms. *Soc Sci Med.* 2011;72:1873–1883.
- [47] Michielsen M, Comijs HC, Semeijn EJ, et al. The comorbidity of anxiety and depressive symptoms in older adults with attention-deficityhyper/activity disorder: a longitudinal study. *J Affect Disord.* 2013;148(2-3):220–227.
- [48] Wunderink L, Nienhuis FJ, Sytema S, et al. Guided discontinuation versus maintenance treatment in remitted first-episode psychosis: relapse rates and functional outcome. *J Clin Psychiatry.* 2007;68(5):654–661.
- [49] Ballard ED, Ionescu DE, Vande Voort JL, et al. Improvement in suicidal ideation after ketamine infusion: relationship to reductions in depression and anxiety. *J Psychiatr Res.* 2014;58:161–166.

- [50] Brunault P, Frammery J, Couet C, et al. Predictors of changes in physical, psychosocial, sexual quality of life, and comfort with food after obesity surgery: a 12-month follow-up study. *Qual Life Res.* 2014;24(2):493–501.
- [51] Seidl H, Hunger M, Leidl R, et al. Cost-effectiveness of nurse-based case management versus usual care for elderly patients with myocardial infarction: results from the korinna study. *Eur J Health Econ.* 2014;16(6):671–681.
- [52] Turner DA, Capuano AW, Wilson SR, et al. Depressive symptoms and cognitive decline in older african americans: two scales and their factors. *Am J Geriatr Psychiatry.* 2014;23(6):568–578.
- [53] Ormel J, Oldehinkel AJ, Sijtsma J, et al. The tracking adolescents' individual lives survey (trails): design, current status, and selected findings. *J Am Acad Child Adolesc Psychiatry.* 2012;51(10):1020–1036.
- [54] Tang L, Song J, Belin TR, et al. A comparison of imputation methods in a longitudinal randomized clinical trial. *Stat Med.* 2005;24:2111–2128.
- [55] Yu LM, Burton A, Rivero Arias O. Evaluation of software for multiple imputation of semi-continuous data. *Stat Method Med Res.* 2007;16:243–258.
- [56] van Ginkel JR, van der Ark R, Sijtsma K. Multiple imputation of item scores in test and questionnaire data, and influence on psychometric results. *Multivariate Behav Res.* 2007;42:387–414.
- [57] Bernaards CA, Belin TR, Schafer JL. Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Stat Med.* 2007;26:1368–1383.
- [58] Lee KJ, Galati JC, Simpson JA, et al. Comparison of methods for imputing ordinal data using multivariate normal imputation: a case study of non-linear effects in a large cohort study. *Stat Med.* 2012;31:4164–4174.
- [59] Horton NJ, Lipsitz SR, Parzen M. A potential for bias when rounding in multiple imputation. *Am Stat.* 2003;57(4):229–232.
- [60] Graham JW, Taylor BJ, Olchowski AE, et al. Planned missing data designs in psychological research. *Multivariate Behav Res.* 2006;11(4):323–243.
- [61] Graham JW, Hofer SM, MacKinnon DP. Maximizing the usefulness of data obtained with planned missing value patterns: an application of maximum likelihood procedures. *Multivariate Behav Res.* 1996;31:197–218.
- [62] Hu F, Zidek JV. Incorporating relevant sample information using the likelihood. Technical Report No. 161. Vancouver, BC: Department of Statistics, The University of British Columbia; 1995.
- [63] Field C, Smith B. Robust estimation: a weighted maximum likelihood approach. *Int Stat Rev.* 1994;62(3):405–424.
- [64] Dupuis DJ, Morgenthaler S. Robust weighted likelihood estimators with an application to bivariate extreme value problems. *Can J Stat.* 2002;30(1):17–36.
- [65] Wang X, van Eeden C, Zidek JV. Asymptotic properties of maximum weighted likelihood estimators. *J Stat Plan Inference.* 2004;119:37–54.
- [66] Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. *Stat Methods Med Res.* 2011;23(3):278–295.
- [67] Vansteelandt S, Carpenter J, Kenward MG. Analysis of incomplete data using inverse probability weighting and doubly robust estimators. *Method: Eur J Res Methods Behav Social Sci.* 2010;6(1):37–48.
- [68] Seaman SR, White IR, Copas AJ, et al. Combining multiple imputation and inverse-probability weighting. *Biometrics.* 2012;68(1):129–137.
- [69] Li L, Shen C, Li X, et al. On weighting approaches for missing data. *Stat Methods Med Res.* 2013;22(1):14–30.
- [70] SAS Institute I. *Sas/stat 9.3 user's guide: The nlmixed procedure.* Cary (NC): SAS Institute Inc.; 2011.

Appendices

Appendix 1. Parameter settings for simulation study

The TRAILS case study was used to generate questionnaire data in our simulation. Thus, we used the set-up and selected parameter values that would be close to the parameter estimates when we would fit model (1). We used $J = 10$ items on 2230 individuals at $T = 4$ time points from TRAILS.

A.1 Full data set: covariates and latent variables

A descriptive analysis of the covariates showed that ages (x_{1t}) had approximately a mean vector $\mu_a^T = (11.1, 13.5, 16.3, 19.0)$ and standard deviation $(0.5, 0.5, 0.7, 0.6)$. The correlation between x_{1t} and $x_{1,t+2}$ was selected $0.6 + 0.1(-1)^T$, for $t = 1$ and 2, and it was approximately equal to 0.6 for any other pair of time points. Therefore, the covariance matrix Σ_a was equal to

$$\Sigma_a = \begin{pmatrix} 0.250 & 0.15 & 0.175 & 0.180 \\ 0.150 & 0.25 & 0.210 & 0.210 \\ 0.175 & 0.21 & 0.490 & 0.252 \\ 0.180 & 0.21 & 0.252 & 0.360 \end{pmatrix}.$$

We selected the success probability p_g as 0.45 to obtain more females than males in the simulation. The mean of the remaining covariates were selected $\mu_c = (0, 0, 0, 0.1)$ with standard deviations $(1, 1, 1, 0.8)$ and correlation matrix

$$\rho_c = \begin{pmatrix} 1 & 0.20 & 0.15 & -0.10 \\ 0.20 & 1 & 0.40 & -0.15 \\ 0.15 & 0.40 & 1 & -0.20 \\ -0.10 & -0.15 & -0.20 & 1 \end{pmatrix},$$

which leads to the variance-covariance matrix as following

$$\Sigma_c = \begin{pmatrix} 1 & 0.20 & 0.15 & -0.08 \\ 0.20 & 1 & 0.40 & -0.12 \\ 0.15 & 0.40 & 1 & -0.16 \\ -0.08 & -0.12 & -0.16 & 0.64 \end{pmatrix}.$$

Then probabilities $p_3, p_4,$ and p_5 were chosen equal to 0.4, 0.1, and 0.2, respectively. Finally, the correlation matrix for the latent variables (Σ_L) was determined as 0.55 for time lags of 1, 0.45 for time lags of 2, and 0.35 for time lags of 3.

A.2 Full data set: items

We analysed the 10 items, from TRAILS, per time point separately using model (1) and in the NLMIXED procedure of SAS [70], version 9.4, to obtain the regression parameters for the covariates, the difficulty and discrimination parameters of each item per time point. Based on the estimates, the coefficient of age (x_{1t}) was determined at -0.1 for the first-seven items and at 0.1 for the remaining items on all time points. For gender (x_2) we used time-varying coefficients, meaning that there was an interaction between time and gender for each item: if Item was 1, 4, 6 or 10, the coefficient was -0.5 for the first visit and -1 for the other visits. For Item 5, 7, and 8, it was -0.5 for the first and last visits and it was -1 for the second and third visits. Finally, the coefficient for Items 2, 3, and 9 were fixed at zero for visit one, -1 for Visit 2 and 3, and -0.5 for Visit 4. We took time stationary coefficients of 0.1 and -0.1 for x_3 and x_4 on all items, respectively. For x_5 , we selected 0.3 for Item 1, 3, 4, 9, and 10; and 0.1 for the rest of the items on all time points. Finally, the regression coefficients for x_6 were selected at -0.1 for item one, three, four, and nine; and at 0.1 for the other items on all time points. Table A1 represents the values for the difficulty and discrimination parameters at each time point.

A.3 Missing indicator variables

For the medium proportion of missing items, the following coefficients were chosen to be implemented in (2). The coefficient for age (x_1) was 0.2 for the second and fourth time points, and 0.1 for the third time point. Effect of gender (x_2) was determined at 0.5 for Item 1, and at 0.2 for the other items at the second visit. It changed to -0.2 and -0.9 for the third and fourth time points, respectively. The effects of x_3 and x_4 were taken zero for all items at all visits. Variable x_5 had an effect of -0.8 for item 1 and -0.4 for the remaining items at visit two, but a constant effect of -0.1 and -0.6 for all items at the third and fourth time points, respectively. The coefficients for the final independent variable x_6 was set to 0.2 for Item 1 and at 0.06 for the other items at the second visit. It was 0.5 for all items at the third and fourth time points. Furthermore, the difficulty parameter was set at zero at all items and time point. The discrimination parameters are listed in Table A2. These setting generated a similar pattern of missingness per item over time as the TRAILS study when we use the independent indicators.

For the small proportion of missing items, we tried to diminish the number of missing items, while still having a dominant influence of the latent variable from the first time point. This changed the difficulty and discrimination parameters to the values listed in Table A2. We used the same coefficient setting variables x_1 , x_2 , and x_5 from the medium setting, but changed the settings for x_3 , x_4 , and x_6 . The coefficients for the 10 items of variable x_3 were selected at (0.1, 0.12, 0.2, 0.1, 0.2, 0.16, 0.9, 0.5, 0.3, 0.4), (0.2, 0.25, 0.18, 0.6, 0.4, 0.24, 1.1, 0.9, 0.7, 0.7), and (0.15, 0.1, 0.3, 0.4, 0.4, 0.2, 1.5, 1.5, 0.4, 0.5) for time points 2, 3, and 4, respectively. For variable x_4 , the coefficients for the second time points were 0.1, 0.25, 0.5, 0.2, 0.3, 0.5, 0.4, 0.8, 0.7, and 0.8 for the 10 items, respectively. For the third time point, the coefficients were 0.3 and 0.4 for Item 1 and Item 2, and 0.2 for the other items.

Table A1. The implemented intercepts and random effect in the simulation.

	a_j				b_j			
	Visit 1	Visit 2	Visit 3	Visit 4	Visit 1	Visit 2	Visit 3	Visit 4
Item 1	-0.5	-1.00	-1.5	-1.5	1.2	1.6	1.6	1.6
Item 2	-1.4	-1.40	-1.9	-2.5	1.6	1.7	1.6	1.7
Item 3	-1.5	-2.00	-2.5	-3.0	1.3	1.4	1.6	1.7
Item 4	-2.5	-3.00	-3.5	-4.0	2.0	2.5	3.0	3.5
Item 5	0.05	0.05	-0.2	-0.5	1.4	1.6	1.6	1.6
Item 6	-1.5	-2.00	-2.5	-3.0	2.0	2.0	2.0	2.0
Item 7	-1.0	-1.50	-2.0	-2.5	1.8	1.8	1.8	1.8
Item 8	-0.6	-0.10	-0.2	-0.5	1.5	1.5	2.0	2.0
Item 9	-3.0	-3.50	-4.0	-4.5	1.5	2.0	1.5	2.0
Item 10	-1.2	-0.20	0.8	-0.2	2.0	2.0	2.0	2.0

Table A2. Intercept and coefficient of the first latent variable in logistic regression model for missing indicator variable.

	Small ($\tilde{a}_{ij}, \tilde{b}_{ij}$)			Medium ($\tilde{a}_{ij} = 0, \tilde{b}_{ij}$)			Large ($\tilde{a}_{ij} = 0, \tilde{b}_{ij}$)		
	Visit 1	Visit 2	Visit 3	Visit 1	Visit 2	Visit 3	Visit 1	Visit 2	Visit 3
Item ₁	(1.5, 1.1)	(1.20, 2.0)	(1.7, 3.2)	1.1	1.5	4.2	2.1	2.5	4.2
Item ₂	(1.1, 1.6)	(1.10, 2.3)	(1.9, 3.3)	0.6	1.8	4.3	1.2	4.6	7.3
Item ₃	(1.2, 1.5)	(0.90, 2.2)	(1.6, 3.3)	0.5	1.7	4.3	1.5	4.7	6.3
Item ₄	(1.4, 1.8)	(1.50, 2.1)	(0.8, 3.4)	0.8	1.8	4.4	3.5	5.8	6.4
Item ₅	(1.1, 1.7)	(1.05, 2.0)	(1.2, 3.2)	0.7	2.0	4.2	1.5	4.5	6.2
Item ₆	(1.1, 1.7)	(1.30, 2.3)	(1.5, 3.2)	0.7	1.9	4.2	2.0	4.9	7.2
Item ₇	(1.9, 1.4)	(1.40, 2.5)	(1.2, 3.3)	0.4	1.8	4.3	1.4	3.6	6.3
Item ₈	(1.6, 1.5)	(1.50, 2.1)	(1.1, 3.2)	0.5	1.8	4.2	1.8	3.0	6.2
Item ₉	(1.3, 1.5)	(1.10, 1.9)	(1.5, 3.4)	0.5	1.4	4.4	1.5	3.4	6.4
Item ₁₀	(1.0, 1.5)	(1.40, 1.7)	(1.1, 3.6)	0.7	1.5	4.6	1.5	3.5	6.6

No effect of x_4 was applied at the fourth visit. The coefficients of x_6 for the 10 items were chosen at the second visit at 0.20, 0.16, 0.19, 0.30, 0.20, 0.16, 0.16, 0.14, 0.20, and 0.60, respectively, and 0.5 for all 10 items at time 3 and 4.

For the large proportion of missingness, we used almost similar settings as the medium setting. The coefficients of covariate age (x_1) at time 2 were 0.2 for Items 1, 3, 4, 8, 9, and Item 10; 0.18 for Item 2 and Item 7; 0.3 for Item 4; and 0.23 for Item 6. For the third time point, it was selected at 0 for Items 1, 2, and 5; -0.01 for Items 3, 7 and 10; -0.03 for Items 4, and 9; and -0.02 for Items 6 and 8. For Visit 4, the coefficients were 0.1 for Items 1 and 2; 0.09 for Items 3, 5, and 9; 0.06 for Item 4; 0.08 for Items 6, 7 and 8 and 0.03 for Item 10. The effect of (x_2) were 0.2 for all items at Visit 2, except for Item 1 which was 0.5. For the third and fourth time points, the coefficients were -0.2 and -0.9 for all items, except for Item 2 which was -2.2 and -2.9 , respectively. The effect of x_3 was fixed at 0.2 and 0.1 for third and fourth visit for Item 2, but it had no effect on the other items at any time points. Variable x_4 had no effect at any items at any time points. The effect of independent variable x_5 was set at -0.8 for Item 1 and at -0.4 for the other items at Visit 2. These coefficients altered to -0.1 and -0.6 for the time 3, and 4, respectively. Covariate x_6 had a coefficient of 0.2 for Item 1 and 0.06 for the other items at Visit 2. At Visit 3 and 4, the effect of x_6 was 1.5 for Item 2 and 0.5 for the other items. Furthermore, the difficulty parameter was set at zero at all items and time point. The discrimination parameters are listed in Table A2.

Appendix 2. SAS syntax

A.4 Code for MI using PMM

```
PROC MI DATA= ITEMLEVEL NIMPUTE=10 OUT=ITEMPMM seed=3478568;
FCS NBITER = 10 REGPMM(ITEM1_2 ITEM2_2 ITEM3_2 ITEM4_2 ITEM5_2
                        ITEM6_2 ITEM7_2 ITEM8_2 ITEM9_2 ITEM10_2
                        ITEM1_3 ITEM2_3 ITEM3_3 ITEM4_3 ITEM5_3
                        ITEM6_3 ITEM7_3 ITEM8_3 ITEM9_3 ITEM10_3
                        ITEM1_4 ITEM2_4 ITEM3_4 ITEM4_4 ITEM5_4
                        ITEM6_4 ITEM7_4 ITEM8_4 ITEM9_4
                        ITEM10_4 / K=3);
VAR ITEM1_1 ITEM2_1 ITEM3_1 ITEM4_1 ITEM5_1
    ITEM6_1 ITEM7_1 ITEM8_1 ITEM9_1 ITEM10_1
    ITEM1_2 ITEM2_2 ITEM3_2 ITEM4_2 ITEM5_2
    ITEM6_2 ITEM7_2 ITEM8_2 ITEM9_2 ITEM10_2
    ITEM1_3 ITEM2_3 ITEM3_3 ITEM4_3 ITEM5_3
    ITEM6_3 ITEM7_3 ITEM8_3 ITEM9_3 ITEM10_3
    ITEM1_4 ITEM2_4 ITEM3_4 ITEM4_4 ITEM5_4
    ITEM6_4 ITEM7_4 ITEM8_4 ITEM9_4 ITEM10_4
    X1 X2 X3 X4 X5 AGE1 AGE2 AGE3 AGE4 ;
RUN;
```

A.5 Code for MI using logistic regression

```
PROC MI DATA = ITEMLEVEL NIMPUTE=10 OUT=ITEMLOG;
CLASS ITEM1_2 ITEM2_2 ITEM3_2 ITEM4_2 ITEM5_2
      ITEM6_2 ITEM7_2 ITEM8_2 ITEM9_2 ITEM10_2
      ITEM1_3 ITEM2_3 ITEM3_3 ITEM4_3 ITEM5_3
      ITEM6_3 ITEM7_3 ITEM8_3 ITEM9_3 ITEM10_3
      ITEM1_4 ITEM2_4 ITEM3_4 ITEM4_4 ITEM5_4
      ITEM6_4 ITEM7_4 ITEM8_4 ITEM9_4 ITEM10_4;
FCS NBITER = 10 LOGISTIC(ITEM1_2 ITEM2_2 ITEM3_2 ITEM4_2 ITEM5_2
                        ITEM6_2 ITEM7_2 ITEM8_2 ITEM9_2 ITEM10_2
                        ITEM1_3 ITEM2_3 ITEM3_3 ITEM4_3 ITEM5_3
```



```
ITEM6_3 ITEM7_3 ITEM8_3 ITEM9_3 ITEM10_3
ITEM1_4 ITEM2_4 ITEM3_4 ITEM4_4 ITEM5_4
ITEM6_4 ITEM7_4 ITEM8_4 ITEM9_4
ITEM10_4);
VAR ITEM1_1 ITEM2_1 ITEM3_1 ITEM4_1 ITEM5_1
ITEM6_1 ITEM7_1 ITEM8_1 ITEM9_1 ITEM10_1
ITEM1_2 ITEM2_2 ITEM3_2 ITEM4_2 ITEM5_2
ITEM6_2 ITEM7_2 ITEM8_2 ITEM9_2 ITEM10_2
ITEM1_3 ITEM2_3 ITEM3_3 ITEM4_3 ITEM5_3
ITEM6_3 ITEM7_3 ITEM8_3 ITEM9_3 ITEM10_3
ITEM1_4 ITEM2_4 ITEM3_4 ITEM4_4 ITEM5_4
ITEM6_4 ITEM7_4 ITEM8_4 ITEM9_4 ITEM10_4
X1 X2 X3 X4 X5 AGE1 AGE2 AGE3 AGE4 ;
RUN;
```