ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2020)
69, *Part* 3, *pp.* 547–564

Multiple imputation of binary multilevel missing not at random data

Angelina Hammon

DIW Berlin, University of Bamberg, and LIfBi, Bamberg, Germany

and Sabine Zinn

DIW Berlin and LIfBi, Bamberg, Germany

[Received February 2019. Final revision January 2020]

Summary. We introduce a selection model-based multilevel imputation approach to be used within the fully conditional specification framework for multiple imputation. Concretely, we apply a censored bivariate probit model to describe binary variables assumed to be missing not at random. The first equation of the model defines the regression model for the missing data mechanism. The second equation specifies the regression model of the variable to be imputed. The non-random selection of the binary data is mapped by correlations between the error terms of the two regression models. Hierarchical data structures are modelled by random intercepts in both equations. To fit the novel imputation model we use maximum likelihood and adaptive Gauss–Hermite quadrature. A comprehensive simulation study shows the overall performance of the approach. We test its usefulness for empirical research by applying it to a common problem in social scientific research: the emergence of educational aspirations. Our software is designed to be used in the R package mice.

Keywords: Fully conditional specification; Missingness not at random; Multilevel data; Multiple imputation; Selection model

1. Introduction

In the social sciences, large-scale surveys with complex data structures have become the norm rather than the exception. Applied statisticians are well aware that, for valid statistical inference, they need to account for the peculiarities arising from complex survey data; see, for example, Gelman *et al.* (1998), Kish and Frankel (1974) and Rubin (1987). Nonetheless, despite urgent appeals for methodologies that are capable of meeting this need, an effective set of survey statistical methods is still lacking. For some time, multiple imputation (MI) (Rubin, 1987) has been the state of the art for handling missing data in surveys. Yet, to the authors' knowledge, no appropriate method exists to handle binary multilevel data under the missingness not at random (MNAR) assumption in the context of MI. Standard applications of MI techniques are usually based on the assumption that the data are missing at random (Rubin, 1976). In many situations, however, it seems entirely realistic to assume that the missing values depend on the incomplete variable Y itself even after conditioning on all the other available data and thus follow an MNAR mechanism. It is well known that this occurs, for example, in variables that are related to income. In this case, usual MI models and implementations may not be sufficient and may result in biased estimates. This paper aims to fill this gap by introducing a selection model-based

Address for correspondence: Angelina Hammon, DIW Berlin, Mohrenstrasse 58, 10117 Berlin, Germany.
E-mail: ahammon@diw.de

 0035–9254/20/69547
Published by John Wiley & Sons Ltd on behalf of the Royal Statistical Society.

MI model to be used within the fully conditional specification (FCS) framework for MI. The FCS–MI approach makes it possible to deal with missing values in variables of several distinct types (Raghunathan *et al.*, 2001; Van Buuren *et al.*, 2006). Our method is an imputation model for binary multilevel data that are assumed to be missing not at random. It is designed to be used in the R package `mice` (Van Buuren and Groothuis-Oudshoorn, 2011). Thus, it is flexible and versatile. A variety of methods exists for addressing multilevel data that are missing at random, e.g. weighting (Asparouhov, 2006; Asparouhov and Muthen, 2006; Pfeffermann *et al.*, 1998), MI (Audigier *et al.*, 2017; Enders *et al.*, 2017; Lüdtke *et al.*, 2017) and likelihood-based methods such as the full information maximum likelihood approach (Larsen, 2011; Wothke, 2000).

Under MNAR the probability of missing data is a function of the unobserved values themselves. As a consequence, each MNAR problem requires specification of the joint distribution $f(Y, R)$ of the complete data Y and the missing data indicator R. Depending on the factorization of this distribution, two kinds of MNAR models result (Little, 2008). Factorizing $f(Y, R)$ as $f(Y)\,f(R|Y)$ gives selection models, and decomposing it as $f(R)\,f(Y|R)$ yields pattern–mixture models. Thus, selection models specify the joint distribution by weighting the marginal distribution of the outcome variable $f(Y)$ by a selection probability $f(R|Y)$, that accounts for non-random non-response and must be modelled explicitly (Rubin, 1974).

In contrast, pattern–mixture models define the full data likelihood as a mixture of distinct response patterns. Thus, different distributions for Y are assumed for units with observed and missing values (Little, 1993; Rubin, 1977). Both model types rely on unverifiable assumptions that cannot be validated by the observed data. In selection models the full data response distribution $f(Y)$ and the conditional distribution of the missing data mechanism $f(R|Y)$ must be specified by underlying parametric assumptions and, in pattern–mixture models, the outcome distribution of the missing values $f(Y_{\text{mis}})$ must be determined by additional external information (Glynn *et al.*, 1986).

It is not possible to test empirically whether the missing data are missing not at random or missing at random since the information that is required to be able to distinguish between these two mechanisms is not available in the data. Furthermore, we must be aware that all types of alternative MNAR models also rely on untestable assumptions. Therefore, it is reasonable to estimate a variety of missing data models with different assumptions, rather than to rely exclusively on one type of model. This non-testability of the assumptions about the missing data mechanism makes sensitivity analysis indispensable for possibly missing not at random data (Molenberghs and Fitzmaurice, 2008). Only in this way can the effect of the assumptions and thus the robustness of statistical inference be assessed. In the context of MI, sensitivity analyses aim to contrast analysis results from missing at random imputed data with those from alternative MNAR models to evaluate whether they lead to different inferences and conclusions. In the existing literature, most methods that apply FCS under MNAR use sensitivity parameters to address distributional differences between the missing and observed units (Hedeker *et al.*, 2007; Resseguier *et al.*, 2011; Tompsett *et al.*, 2018; Van Buuren *et al.*, 1999). Such parameters cannot be estimated from observed data. Instead, plausible value ranges must be provided by experts in the respective research field. In the social sciences, it is usually difficult to define such value ranges because of the widespread lack of objective and generally valid comparative figures. We therefore present an MNAR imputation model that does not include sensitivity parameters but is completely identified by its distributional assumptions.

The idea of using a selection model (Rubin, 1974) in the context of FCS–MI is not new. Galimard *et al.* (2016) used a two-stage selection model for imputing continuous missing not at random data. Very recently, Galimard *et al.* (2015, 2018) presented a variant of their model that applies to binary missing not at random data. Their basic idea is to use a bivariate probit

model to specify the response probability and the binary outcome of a single-level model jointly. We extend their approach by adding random intercepts to the selection and outcome model to account for potential multilevel structures in the data. We use a maximum likelihood approach and adaptive Gauss–Hermite quadrature (AGHQ) procedure to fit our imputation model. The derivation of the formulae that is used will be described in detail below.

Most multilevel MNAR approaches originate from biostatistics and deal with continuous multilevel data; for comprehensive overviews, see, for example, Little (2008), Molenberghs and Fitzmaurice (2008) and Molenberghs *et al.* (2008). The lack of related applications in the social sciences, where variables are mostly ordinal and binary, is striking. We make a first step towards rectifying this situation by applying our novel approach to a common research problem in the social (educational) sciences. Before doing so, we examine the overall performance of our new approach by conducting a comprehensive simulation study. By means of standard errors, relative bias and coverage rates, we compare the performance of our novel MI–FCS MNAR imputation model with other common imputation strategies that cope either with binary multilevel data or with binary missing not at random data, but never with both simultaneously. Furthermore, we test the robustness of our approach under misspecified missing data models. We then apply our novel method to analyse the effect of personal attributions and social background factors on the educational aspirations of ninth-grade students in Germany.

It is also advisable to compare the performance of our method with that of alternative MNAR models such as pattern–mixture models. However, in the context of FCS–MI, no such method currently exists. The development of such a method is beyond the scope of this paper and thus is left for future work.

The remainder of this paper is structured as follows. First, we describe the new imputation method and the underlying model. This is followed by the simulation study and the presentation of its results. Thereafter, we apply the approach to empirical data from the German National Educational Panel Study (NEPS). We conclude with a short summary of the results, a discussion of some critical issues and tasks for future work.

2. Method

The basic idea of FCS–MI is to specify separate imputation models for each incomplete variable and to impute the missing data variable by variable, i.e. for a binary variable with missing values a model describing this variable appropriately is required. If data are additionally missing not at random, then the mechanism that caused the missing values must also be modelled. For this, like Galimard *et al.* (2015, 2016, 2018) we use a selection-model-based approach. Combined with the binary variable Y to be imputed, this yields a two-equation system: one equation for the selection process and one equation describing Y. We use a bivariate probit model with sample selection (Greene, 2012; Wooldridge, 2002), i.e. a censored bivariate probit model, to specify this two-equation system. Multilevel structures in the data are accounted for by expanding the bivariate probit model by a random-intercept term. This expanded bivariate model serves as an imputation model to impute missing values of Y within the iterative FCS–MI scheme. In what follows, we describe this model in detail and present an efficient way to estimate it. We then present the imputation algorithm that will be used to obtain plausible replacements for the missing values of Y. Here, R describes the missing data indicator of Y that takes the value 1 if Y is observed and 0 otherwise. Observations of Y and R are denoted by y and r.

2.1. Imputation model: censored bivariate probit model with random intercept

Assume that the data at hand contain $j = 1, \ldots, J$ clusters consisting of $i = 1, \ldots, n_j$ individuals

respectively. By using the standard probit specification based on a latent variable formulation, the model can be specified as follows:

$$\begin{aligned} r^*_{ji} &= \beta_R x_{R,ji} + \alpha_{R,j} + \epsilon_{R,ji}, \\ y^*_{ji} &= \beta_Y x_{Y,ji} + \alpha_{Y,j} + \epsilon_{Y,ji} \end{aligned} \tag{2.1}$$

with

$$r_{ji} = \mathbf{1}(r^*_{ji} > 0),$$

$$y_{ji} = \begin{cases} 1 & \text{if } (y^*_{ji} > 0 \,\&\, r_{ji} = 1), \\ 0 & \text{if } (y^*_{ji} < 0 \,\&\, r_{ji} = 1), \\ \text{not applicable} & \text{if } r_{ji} = 0. \end{cases}$$

The first equation accounts for the non-random selection process, i.e. for the missing data mechanism in our case. The second equation models the focal variable Y and defines the outcome equation. The asterisk denotes the latent variables r^*_{ji} and y^*_{ji}, whose observed equivalents are r_{ji} and y_{ji}. The covariates of the two regression equations are x_R and x_Y, and β_R and β_Y are the related coefficients. $\alpha_{R,j}$ and $\alpha_{Y,j}$ are the random intercepts for describing cluster effects, and $\epsilon_{R,ji}$ and $\epsilon_{Y,ji}$ are the error terms. The function $\mathbf{1}$ denotes the indicator function and 'not applicable' denotes a missing value. To assure model identifiability, x_Y must be a subset of x_R and $x^e_R = x_R \setminus x_Y$ to be highly correlated with r and hardly connected to y (Rendtel, 1992). The set x^e_R is called the exclusion restriction. The selection and the outcome equation are linked through correlated error terms and random intercepts:

$$\begin{aligned} \begin{pmatrix} \epsilon_R \\ \epsilon_Y \end{pmatrix} &\sim N\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right\}, \\ \begin{pmatrix} \alpha_R \\ \alpha_Y \end{pmatrix} &\sim N\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_R & \tau\sigma_R\sigma_Y \\ \tau\sigma_R\sigma_Y & \sigma^2_Y \end{pmatrix} \right\}. \end{aligned} \tag{2.2}$$

Here ρ describes the correlation of the bivariate distribution of R^* and Y^*, and therefore models the relationship between the selection and outcome equation. Clearly, this two-equation system specifies only the dependence of the missing-data mechanism on the outcome variable appropriately if the normality assumptions hold. To capture potential dependences of the missing data mechanism on the cluster structure of the data, the random intercepts α_R and α_Y are also allowed to depend on each other. τ denotes the correlation of the bivariate normal distribution of α_R and α_Y, and $\boldsymbol{\Sigma}$ their variance–covariance matrix. In this paper, only a two-level hierarchy is considered, but the extension to further levels is straightforward.

The log-likelihood function of the two-equation model (2.1) can be expressed as (see, for example, Greene (2012))

$$\begin{aligned} \ln(L) = \sum_{j=1}^{J} \ln\Bigg(\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \prod_{i=1}^{n_j} [& r_{ji} y_{ji} \Phi_2(\beta_R x_{R,ji} + \alpha_{R,j}, \beta_Y x_{Y,ji} + \alpha_{Y,j}, \rho) \\ & + r_{ji}(1 - y_{ji})\, \Phi_2\{\beta_R x_{R,ji} + \alpha_{R,j}, -(\beta_Y x_{Y,ji} + \alpha_{Y,j}), -\rho\} \\ & + (1 - r_{ji})\, \Phi\{-(\beta_R x_{R,ji} + \alpha_{R,j})\}] \, \phi_2(\alpha_{R,j}, \alpha_{Y,j} | \mathbf{0}, \boldsymbol{\Sigma}) \Bigg) \mathrm{d}\alpha_{R,j} \mathrm{d}\alpha_{Y,j}. \end{aligned} \tag{2.3}$$

Here, $\Phi_2(\cdot)$ denotes the cumulative distribution function of the bivariate standard normal distribution and $\Phi(\cdot)$ is the cumulative distribution function of the univariate standard normal. The function $\phi_2(\cdot|\mathbf{0}, \boldsymbol{\Sigma})$ is the probability density function of a bivariate normal distribution with mean 0 and variance–covariance matrix $\boldsymbol{\Sigma}$. The two random intercepts α_R and α_Y are

treated as nuisance parameters and are integrated out. This works in the same way as for the univariate panel model with one random intercept (see, for example, Butler and Moffitt (1982)). The double integral of the log-likelihood function (2.3) has no closed form solution. One way to solve the integral nevertheless is to approximate the area under the integrand. This can be done by using either simulation (e.g. Cappellari and Jenkins (2003), Greene (2004) and Train (2009)) or quadrature techniques (e.g. Delattre and Moussa (2015) and Mulkay (2015)). In this paper, we use Gauss–Hermite quadrature (GHQ) to solve the double integral. The reason is that simulation is expected to take much more computational time for the two dimensions that we have than quadrature, which is of tremendous disadvantage in the context of MI when data are imputed several times and usually also for several variables. Simulation only becomes beneficial with higher dimensions since the required number of iterations for approximating the integrals does not depend on the number of dimensions. GHQ approximates an integral of a specific form by a weighted sum of the integrand evaluated at predetermined abscissas of the variable that is integrated out (see Abramowitz and Stegun (1964) and Davis and Rabinowitz (1967)). These predetermined abscissas are called quadrature points. For example, in the one-dimensional case the integral of the form

$$I = \int_{-\infty}^{\infty} f(x) \exp(-x^2) \mathrm{d}x$$

can be approximated by

$$\tilde{I} = \sum_{p=1}^{P} \omega_p f(a_p),$$

where the quadrature points a_p are set to be the nodes of the Hermite polynomial and ω_p are the corresponding weights with $p = 1, \ldots, P$. By design, the quadrature points are set symmetrically around zero. The accuracy of the Gauss–Hermite approximation $\tilde{I}$ depends on the chosen number P of quadrature points. Ideally, P is determined by investigating the convergence behaviour of $\tilde{I}$ when P is increased. However, Lesaffre and Spiessens (2001) showed that a number of $P = 10$ is often sufficient and differences by further increasing P are only minimal. The Gauss–Hermite weights and nodes can be found in the tables of Abramowitz and Stegun (1964) or can be computed by using an algorithm that was proposed by Golub and Welsch (1969). An improved version of GHQ is AGHQ (Liu and Donald, 1994; Naylor and Smith, 1982). Here, in contrast with traditional GHQ, the quadrature points are set symmetrically around the maximum value of the integrand. In other words, AGHQ shifts and scales the quadrature locations to place them under the peak of the integrand, so that the function is evaluated where the area is expected to be largest. Applying the AGHQ approach to log-likelihood equation (2.3) gives the following approximation (the corresponding calculation steps yielding this formula are given in the on-line supplementary material provided along with this paper):

$$\begin{aligned}
\ln(L) \simeq \sum_{j=1}^{J} \ln\Bigg(& |\boldsymbol{\Omega}_j|^{1/2} \times 2 \sum_{p_1=1}^{P} \sum_{p_2=1}^{P} \omega_{p_1} \omega_{p_2} \prod_{i=1}^{n_j} [r_{ji} y_{ji} \Phi_2(\beta_R x_{R,ji} + \tilde{a}_{jp_1}, \beta_Y x_{Y,ji} + \tilde{a}_{jp_2}, \rho) \\
& + r_{ji}(1 - y_{ji}) \Phi_2\{\beta_R x_{R,ji} + \tilde{a}_{jp_1}, -(\beta_Y x_{Y,ji} + \tilde{a}_{jp_2}), -\rho\} \\
& + (1 - r_{ji})\, \Phi\{-(\beta_R x_{R,ji} + \tilde{a}_{jp_1})\}] \phi_2(\tilde{\mathbf{a}}_{jp} | \mathbf{0}, \boldsymbol{\Sigma}) \exp(\mathbf{a}_p' \mathbf{a}_p) \Bigg),
\end{aligned} \tag{2.4}$$

where $p_1 = 1, \ldots, P$ and $p_2 = 1, \ldots, P$ are the quadrature points for the selection equation and for the outcome equation respectively. In contrast with GHQ, AGHQ enables us to specify quadrature points for each cluster separately. It can be expected that this improves the approximation of the cluster-specific integrals $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [\cdots] \phi_2(\cdots) \mathrm{d}\alpha_{R,j} \mathrm{d}\alpha_{Y,j}$ (compare log-likelihood (2.3)).

The related bivariate quadrature points $\tilde{\mathbf{a}}_{jp} = (\tilde{a}_{jp_1}\tilde{a}_{jp_2})'$ with $p = (p_1, p_2)'$ are defined as

$$\tilde{\mathbf{a}}_{jp} = \boldsymbol{\mu}_j + \sqrt{2}\boldsymbol{\Omega}_j^{1/2}\mathbf{a}_p',$$

where $\mathbf{a}_p = (a_{p_1}, a_{p_2})'$ and $\boldsymbol{\omega}_p = (\omega_{p_1}\omega_{p_2})'$ are the standard Gauss–Hermite nodes and weights. Here, the matrix $\boldsymbol{\Omega}_j$ scales $\mathbf{a}_p$ and the vector $\boldsymbol{\mu}_j$ centres them. The function $|\boldsymbol{\Omega}_j|$ denotes the determinant of $\boldsymbol{\Omega}_j$. The square root of $\boldsymbol{\Omega}_j$, $\boldsymbol{\Omega}_j^{1/2}$, can be properly described by the lower triangular matrix $\mathbf{T}$ of the Cholesky decomposition of $\boldsymbol{\Omega}_j = \mathbf{T}\mathbf{T}'$.

There are different approaches to specify $\boldsymbol{\mu}_j$ and $\boldsymbol{\Omega}_j$. Liu and Donald (1994) recommended centring the nodes with respect to the mode of the integrand and scaling them according to the negative inverse Hessian matrix (curvature) at the mode. In the case considered, the mode is the most likely value for the random effects given the observed data and the current estimates of all the other model parameters. The integrand of the likelihood l_j for cluster j, that is necessary to calculate the cluster-specific mode and curvature, is

$$\begin{aligned} l_j = \prod_{i=1}^{n_j} [& r_{ji} y_{ji} \Phi_2(\beta_R x_{R,ji} + \alpha_{R,j}, \beta_Y x_{Y,ji} + \alpha_{Y,j}, \rho) + r_{ji}(1 - y_{ji})\Phi_2\{\beta_R x_{R,ji} + \alpha_{R,j}, \\ & -(\beta_Y x_{Y,ji} + \alpha_{Y,j}), -\rho\} + (1 - r_{ji})\Phi\{-(\beta_R x_{R,ji} + \alpha_{R,j})\}]\phi_2(\alpha_{R,j}, \alpha_{Y,j} | \mathbf{0}, \boldsymbol{\Sigma}). \end{aligned}$$

Estimators $\hat{\boldsymbol{\mu}}_j = (\hat{\mu}_{R,j}, \hat{\mu}_{Y,j})'$ of the modes $\boldsymbol{\mu}_j = (\mu_{R,j}, \mu_{Y,j})'$ of the two random intercepts for cluster j can be computed, for example, by

$$\hat{\boldsymbol{\mu}}_j = \underset{(\alpha_{R,j}, \alpha_{Y,j})}{\arg\max} \ln(l_j).$$

The related curvature matrix at the modes, $\hat{\boldsymbol{\Omega}}_j$, is a proper estimator for $\boldsymbol{\Omega}_j$. It is defined as

$$\hat{\boldsymbol{\Omega}}_j = \begin{pmatrix} -\dfrac{\partial^2 l_j}{\partial \alpha_{R,j}^2} & -\dfrac{\partial^2 l_j}{\partial \alpha_{R,j}\alpha_{Y,j}} \\ -\dfrac{\partial^2 l_j}{\partial \alpha_{R,j}\alpha_{Y,j}} & -\dfrac{\partial^2 l_j}{\partial \alpha_{Y,j}^2} \end{pmatrix}^{-1}.$$

The estimator $\hat{\boldsymbol{\mu}}_j$ must be found by some numerical optimization algorithm such as the Nelder–Mead method, whereas $\hat{\boldsymbol{\Omega}}_j$ can be calculated analytically or also solved numerically. In general, AGHQ is clearly superior to ordinary GHQ since it dramatically reduces the number of necessary quadrature points to approximate a given integral by sampling the nodes in the relevant region of the function. Although additional time is needed to compute the mode and curvature at each maximization iteration, many fewer quadrature points are required for the same approximation accuracy (Lesaffre and Spiessens, 2001). This applies especially when the mode of the integrand is far from 0. We rely on the standard maximum likelihood approach to estimate the parameters of the approximated log-likelihood function (2.4). For numerical optimization we suggest using the Broyden–Fletcher–Goldfarb–Shanno method (e.g. Goldfarb (1970))—a very powerful and efficient optimization algorithm for solving unconstrained non-linear optimization problems that belongs to the group of quasi-Newton methods. These methods do not require the computation of the Hessian matrix but approximate it in each iteration by using the gradients. This fact makes them computationally very attractive (e.g. Nocedal and Wright (2006)). To speed up the maximization process of parameter estimation, we calculated the analytic gradients of equation (2.4), which can be found in the on-line supplementary material, and use them during optimization. Note that, during the optimization procedure, $\hat{\boldsymbol{\mu}}_j$ and $\hat{\boldsymbol{\Omega}}_j$ must be computed for each individual cluster in each iteration step.

2.2. Imputation algorithm: fully conditional specification

With the imputation model at hand and an adequate method to estimate it, missing values can now be imputed. As already stated, in FCS–MI, the incomplete variables are imputed sequentially in a univariate fashion. For this, plausible replacements are drawn variable by variable from the related conditional densities. In our case, the bivariate probit model (2.1) determines the conditional density of Y. Let $\theta = (\beta_Y, \beta_R, r, \xi_Y^2, \xi_R^2, z)$ be the unknown parameters of the bivariate probit model, where $r = \tanh^{-1}(\rho)$, $\xi_Y^2 = \ln(\sigma_Y^2)$, $\xi_R^2 = \ln(\sigma_R^2)$ and $z = \tanh^{-1}(\tau)$ are common transformations to preserve the range constraints of the parameters during maximization. At each iteration the following five steps are conducted to impute the missing values of Y. At each FCS–MI iteration step the approximated log-likelihood (2.4) must be maximized to obtain updated estimates $\hat{\theta}$ for θ. Furthermore, to ensure a proper imputation procedure, parameter uncertainty must be considered (Rubin, 1987). This is achieved by drawing parameter candidates $\dot{\theta}$ by using a normal approximation to the posterior distribution of $\hat{\theta}$ (e.g. Gelman *et al.* (2013), chapter 4).

Step 1: estimate model parameters by maximum likelihood and AGHQ by using equation (2.4), which yields

(a) $\hat{\theta} = (\hat{\beta}_Y, \hat{\beta}_R, \hat{r}, \hat{\xi}_Y^2, \hat{\xi}_R^2, \hat{z})$ and
(b) $\hat{\psi}$, the variance–covariance matrix of $\hat{\theta}$.

Step 2: draw $\dot{\theta} = (\dot{\beta}_Y, \dot{\beta}_R, \dot{r}, \dot{\xi}_Y^2, \dot{\xi}_R^2, \dot{z})$ from $N(\hat{\theta}, \hat{\psi})$, and retransform $\dot{\rho} = \tanh(\dot{r})$, $\dot{\tau} = \tanh(\dot{z})$, and $\dot{\sigma}_Y^2 = \exp(\dot{\xi}_Y^2)$ and $\dot{\sigma}_R^2 = \exp(\dot{\xi}_R^2)$.
Step 3: draw random-intercept candidates $(\dot{\alpha}_{R,j}, \dot{\alpha}_{Y,j})'$ for each cluster j from $N(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Omega}}_j)$.
Step 4: calculate for each unit with missing Y the probability $\dot{p}$ that Y equals 1:

$$\dot{p} = P(Y = 1 | X_Y, X_R, R = 0) = \frac{\Phi_2\{X_Y\dot{\beta}_Y + \dot{\alpha}_Y, -(X_R\dot{\beta}_R + \dot{\alpha}_R), -\dot{\rho}\}}{\Phi\{-(X_R\dot{\beta}_R + \dot{\alpha}_R)\}}.$$

Step 5: draw for each missing value Y_{mis} a replacement from a Bernoulli distribution with success probability $\dot{p}$.

To generate M imputed data sets, these steps are repeated M times.

This imputation algorithm extends the work of Galimard *et al.* (2018). We have implemented it in a way that the algorithm can be used within the `mice()` function of the R package `mice`. (The corresponding source code is available from `http://github.com/AngelinaHammon/PaperBinaryMNARmultilevelData`.) In FCS–MI, it is common practice that each incomplete variable is also a potential predictor in the imputation models for all the other variables. This applies to Y as well. Since by assumption the missing data indicators R of Y and X_Y are correlated (determined by the selection equation), R must be included as predictor in the imputation models of all of the other incomplete variables that are part of X_Y; see also Galimard *et al.* (2016). Otherwise biased imputations may arise.

3. Simulation study

To evaluate the performance of our novel imputation procedure, we conduct a set of distinct Monte Carlo simulation studies, using different data-generating processes to represent possible real world scenarios. For clarity, we concentrate on the univariate imputation model of Y, and we assume that all the covariates considered are observed completely. However, as already stated, an application of the algorithm to multivariate missing data is straightforward. The number of replications is set to 1000.

3.1. Data generation

In sum, we consider five simulation scenarios. The total sample size is set to $n = 2500$ and the number of clusters equals $m = 50$, leading to a cluster size of $n_j = 50$, $j = 1, \ldots, m$. These numbers are in line with cluster and sample sizes that are commonly used in educational research, for example, when describing students in schools. For simplicity, we assume that all clusters comprise the same number of units. However, the method can also be applied without any problems in case of different cluster sizes. In any simulation scenario, we initially generate complete data sets with one binary outcome variable y_{ji}, $i = 1, \ldots, n_j$, and three different normally distributed covariates $x_{1,ji}, x_{2,ji}$ and $x_{3,ji}$ according to

$$x_{1,ji} \sim N(0, 0.3^2),$$
$$x_{2,ji} \sim N(0, 0.8^2),$$
$$x_{3,ji} \sim N(0, 4^2)$$

and

$$y^*_{ji} = 0.25 + x_{1,ji} + 0.5x_{2,ji} + \alpha_{Y,j} + \epsilon_{Y,ji} \qquad y_{ji} = \mathbf{1}(y^*_{ji} > 0).$$

Here $\alpha_{Y,j}$ and $\epsilon_{Y,ji}$ are drawn according to the model assumptions (2.2) with $\sigma^2_R = 0.5$ and $\sigma^2_Y = 0.9$. This yields an intraclass correlation of about 0.3 for the selection indicator r and of approximately 0.45 for the outcome variable y. Missing values are imposed on y_{ji} by specifying a model for the response indicator r_{ji}, where r_{ji} equals 1 if y_{ji} is observed and is 0 otherwise. To assess the performance of our imputation method under distinct (realistic) missing data situations, we implement models for five different missing data mechanisms. We specify four models for MNAR and one model for missingness at random (MAR). Depending on the mechanism that is considered the parameters ρ and τ of equation (2.2) take varying values expressing different relationships between the response indicator r and the outcome variable y. We include different types of MNAR missing data, where we assume that the probability of observing y_{ji} increases with the value of y^*_{ji}. Under the first three MNAR scenarios (MNAR selection), missing data are produced by using the following parameterization of the selection equation:

$$r^*_{ji} = 0.5 + 1.5x_{1,ji} - 0.25x_{2,ji} + 0.1x_{3,ji} + \alpha_{R,j} + \epsilon_{R,ji} \qquad r_{ji} = \mathbf{1}(r^*_{ji} > 0). \tag{3.1}$$

To take into account different magnitudes of correlation between y_{ji} and r_{ji}, we use three values for ρ, namely $\rho \in \{0.3, 0.6, 0.9\}$, reflecting weak, medium and strong correlation. We set $\tau = 0.5$ to allow for a medium correlation between the random intercepts of both equations. The variable x_3 represents the exclusion criterion. To evaluate our method also in an MNAR situation, where the missing data mechanism does not strictly follow the selection model specification of the imputation procedure (MNAR non-selection), we consider another MNAR scenario, where the missing data are imposed by

$$P(r_{ji} = 1) = \Phi(0.8 + 1.75y^*_{ji} + 1.5x_{1,ji} - 2.5x_{2,ji} + \alpha_{R,j}) \qquad r_{ji} \sim \text{Ber}\{P(r_{ji} = 1)\}.$$

Here, $\text{Ber}(\cdots)$ denotes the Bernoulli distribution, and ρ and τ of equation (2.2) are set to 0. Since it is not possible to test empirically whether the missing data mechanism at hand is MAR or MNAR, sensitivity analyses incorporating alternative imputation models with additional external assumptions are the only means of detecting a potential MNAR mechanism. The underlying idea is that, in the case of MAR the inferences should not differ between the distinct MAR and MNAR imputation methods. Thus, to conduct effective sensitivity analyses it is crucial that the alternative imputation models can not only handle data missing not at random but also yield valid inferences under MAR. Therefore, we additionally consider an MAR scenario where the

missingness does not depend on y_{ji} to evaluate how our new method performs under MAR. For this, we specify the latent response indicator r^*_{ji} by using equation (3.1) with $\rho = 0$ and $\tau = 0$. All missing data scenarios that were examined yield approximately 35% missing values in y. The complete code for data generation and analysis of our simulation study is available from `http://github.com/AngelinaHammon/PaperBinaryMNARmultilevelData`.

3.2. Data analysis

To assess the adequacy of our new imputation method with AGHQ (referred to in what follows as MNAR AGHQ), its performance will be compared with that of other relevant imputation procedures including methods that assume MAR. We also tested the performance of our method by using the GHQ approximation. However, the corresponding results are slightly worse than under MNAR AGHQ. Thus, they are not reported. As MAR imputation methods, we use a mixed effects logistic regression as imputation approach (MAR mixed) (Zinn, 2013) as well as a technique that uses a two-stage estimation approach (MAR 2-stage) (Resche-Rigon and White, 2018). Besides these methods, which take into account the multilevel structure of the data, we also evaluate the procedure of Galimard *et al.* (2018, 2015) that handles MNAR by using a selection model approach but is only designed for single-level data (MNAR Galimard). As a benchmark scenario, we present the results of a complete-case analysis (CCA), which in the case at hand (only missing values in y) is also valid under MAR (Von Hippel, 2007). We used $M = 10$ imputations for each scenario and imputation procedure. Since here we focus on only univariate missing data, which are a special case of monotone missingness, there is no need to iterate the `mice` algorithm.

Each completed data set is analysed by estimating a mixed effects probit regression on y with covariates x_1 and x_2. For this, we use the `glmer()` function of the R package `lme4` (Bates *et al.*, 2015). After estimation, all the estimates are pooled by using Rubin's combining rules (Rubin, 1987). We assess the performance of each imputation method by using the empirical means of the parameter estimates, their relative bias and the empirical standard errors of the estimates, as well as the root mean square of the estimated standard errors. Furthermore, we derive and evaluate the coverage rates of the nominal 95% confidence intervals.

3.3. Results

For the various imputation strategies and simulation scenarios including the MNAR scenario based on a selection model with medium correlation, i.e. for $\rho = 0.6$, Table 1 shows the results for the regression parameter β_1 of the first covariate x_1. Table 2 gives the estimates for the slope parameter β_2 of variable x_2. The results for the selection-model-based scenarios with low and high correlation, i.e. $\rho \in \{0.3, 0.9\}$, are not reported here since they are similar in terms of relative bias and coverage rates. However, they can be found in the on-line supplementary material to this paper. If the true missing data mechanism is MAR, the CCA and the imputation model based on a mixed logistic regression (MAR mixed), which are both designed for this type of missing data, perform—as expected—very well in terms of bias. Surprisingly, the MAR 2-stage approach performs rather badly under MAR with regard to both bias and coverage. Apparently, the two-model strategy is not suitable for the kind of hierarchical data situation that is considered in our simulation study. The results of Audigier *et al.* (2017) suggest that the method introduces bias with cluster sizes below 100, which might be the reason for its poor performance in our simulation study. Our novel approach MNAR AGHQ performs well under the MAR scenario, with an average relative downward bias of 1.98% and a reasonable coverage rate of 96% for β_1. The bias is slightly higher than for the CCA or MAR mixed, but nevertheless these results confirm

Table 1. Simulation results for $\beta_1 = 1$ estimates (with $\tau = 0.5$ and $\rho = 0.6$ for MNAR selection) in 1000 simulation runs†

Method	*Mechanism*	*emp.mean*	*rel.bias (%)*	SE_{emp}	SE_{mod}	*CR (%)*
Before deletion	MAR	1.0020	0.51	0.1032	0.0983	95.6
	MNAR selection	1.0038	0.43	0.1030	0.1045	93.2
	MNAR non-selection	1.0055	0.51	0.1033	0.1090	94.4
CCA	MAR	1.0064	0.95	0.1339	0.1291	93.6
	MNAR selection	0.6765	−32.31	0.1368	0.1282	38.0
	MNAR non-selection	0.5465	−45.37	0.1418	0.1843	16.8
MAR mixed	MAR	1.0023	0.54	0.1499	0.1332	96.4
	MNAR selection	0.6524	−34.72	0.1545	0.1343	42.0
	MNAR non-selection	0.5462	−45.40	0.1519	0.1815	18.0
MAR 2-stage	MAR	0.8931	−10.42	0.1404	0.1232	92.0
	MNAR selection	0.6146	−38.50	0.1438	0.1238	21.2
	MNAR non-selection	0.4855	−51.46	0.1481	0.1622	8.8
MNAR AGHQ	MAR	1.0166	−1.98	0.1840	0.1730	96.0
	MNAR selection	0.9900	−0.94	0.1560	0.1476	95.2
	MNAR non-selection	0.9914	−0.89	0.1396	0.1291	96.8
MNAR Galimard	MAR	0.8342	−16.32	0.1815	0.1708	86.8
	MNAR selection	0.8322	−16.73	0.1581	0.1466	85.2
	MNAR non-selection	0.8772	−12.31	0.1435	0.1296	88.8

†emp.mean denotes the empirical mean of the estimates, rel.bias the relative bias, SE_{emp} the observed standard errors across all simulations, SE_{mod} the root mean square of the estimated standard errors and CR the nominal coverage rate. The Monte Carlo standard error of the coverage rate is 0.69% and the maximum Monte Carlo standard error of the bias is 0.58%. The formulae to compute these quantities have been taken from Morris *et al.* (2019).

Table 2. Simulation results for $\beta_2 = 0.5$ estimates (with $\tau = 0.5$ and $\rho = 0.6$ for MNAR selection) in 1000 simulation runs†

Method	*Mechanism*	*emp.mean*	*rel.bias (%)*	SE_{emp}	SE_{mod}	*CR (%)*
Before deletion	MAR	0.5019	0.70	0.0397	0.0405	96.4
	MNAR selection	0.4969	−0.72	0.0395	0.0378	96.0
	MNAR non-selection	0.5017	0.19	0.0397	0.0390	96.4
CCA	MAR	0.5031	0.93	0.0502	0.0519	93.6
	MNAR selection	0.6050	20.87	0.0532	0.0517	47.2
	MNAR non-selection	0.9706	93.84	0.0662	0.0625	0.0
MAR mixed	MAR	0.5004	0.39	0.0524	0.0522	94.8
	MNAR selection	0.6061	21.10	0.0552	0.0523	55.2
	MNAR non-selection	0.9622	92.18	0.0744	0.0645	0.0
MAR 2-stage	MAR	0.4146	−16.81	0.0518	0.0495	66.0
	MNAR selection	0.5006	0.01	0.0558	0.0477	97.2
	MNAR non-selection	0.8286	65.49	0.0750	0.0804	4.0
MNAR AGHQ	MAR	0.4916	−1.29	0.0574	0.0610	91.2
	MNAR selection	0.4882	2.47	0.0693	0.0653	92.8
	MNAR non-selection	0.5183	3.52	0.1068	0.0992	95.2
MNAR Galimard	MAR	0.4083	−18.09	0.0561	0.0558	59.6
	MNAR selection	0.4141	−17.26	0.0607	0.0621	70.8
	MNAR non-selection	0.4457	−10.99	0.1044	0.0942	90.0

†emp.mean denotes the empirical mean of the estimates, rel.bias the relative bias, SE_{emp} the observed standard errors across all simulations, SE_{mod} the root mean square of the estimated standard errors and CR the nominal coverage rate. The Monte Carlo standard error of the coverage rate is 0.69% and the maximum Monte Carlo standard error of the bias is 0.34%. The formulae to compute these quantities have been taken from Morris *et al.* (2019).

that our novel approach also works for missing data that are in fact missing at random, which is a crucial property for conducting adequate sensitivity analyses. In the considered MNAR scenarios, the MNAR AGHQ method clearly outperforms all competing approaches. For β_1, it yields, under both MNAR conditions, a relative bias of lower than 1% and coverage rates near the nominal coverage probability. The three MAR methods underestimate β_1 up to 51.46% in both MNAR scenarios. The approach of Galimard (MNAR Galimard) yields biased estimates for β_1 in all scenarios. This is caused by the fact that Galimard's model does not induce any multilevel structure into the imputed values.

In principle, the results for parameter β_2 are similar to those of parameter β_1. The MAR mixed approach shows a high upward bias in all the considered MNAR scenarios along with very low coverage rates. Interestingly, the MAR 2-stage technique leads to an unbiased estimate for β_2 and to a nominal coverage of 97.2% in the MNAR scenario based on the selection model and even outperforms MNAR AGHQ, which shows a relative bias of 2.47% and a coverage rate of 92.8%. In contrast, MAR 2-stage again yields a poor coverage rate and biased estimate in the presence of MAR. All three MAR methods overestimate β_2 up to 93.84% under MNAR. For β_2, our new approach MNAR AGHQ again shows reasonable performance in terms of bias and coverage in all the scenarios considered. However, in the data situation where missing data are created under a non-selection model, the estimate is slightly more biased than for the other scenarios. In addition, the bias is also higher than for the estimate of β_1 in the same missing data situation. Nevertheless, the average relative bias of 3.52% still lies within an acceptable range, in particular compared with the performance of the other methods investigated, which clearly underperform relative to MNAR AGHQ. Under MAR, MNAR AGHQ shows a slightly lower coverage rate for β_2 than expected, but it still lies in a reasonable range. As for β_1 the MNAR Galimard model biases the estimates of β_2 downwards and results in too low coverage rates. Hence, even if the MNAR mechanism is modelled, the failure to consider the hierarchical data structure during imputation leads to incorrect estimates of fixed effects parameters.

4. Analysis of educational aspirations

We apply our novel imputation method to a frequently studied problem in educational research: young people's educational aspirations and the effect of their social background. Our analysis focuses on ninth-grade students attending lower secondary school, *Hauptschule*, the lowest track of secondary school in Germany. This group is particularly affected by social disadvantage (Schneider, 2018; Wößmann, 2007) and thus is of special interest to educational researchers. This is especially true for the relationship between parental education and the degree that students aspire to obtain. We use wave 1 of the NEPS starting cohort 'School and vocational training: educational pathways of students in grade 9 and higher' to study the aspirations of ninth graders to graduate with a degree that is higher than the degree that is offered by the school that they are currently visiting. For students attending *Hauptschule*, this is either an intermediate secondary degree or a degree allowing for university admission. We study the effect of parental education on a student's aspirations by using the information on whether a student's mother has a university admission certificate (UAC) or not. As personal attributes that may influence students' aspirations we consider their grades in mathematics and German and their competencies in mathematics and reading, and their gender, as well as their migration background (measured by generation status smaller than 3.5). Competency scores are estimated as weighted maximum likelihood estimates (Warm, 1989). Detailed information on the measurement of competencies in the NEPS is given in Duchhardt and Gerdes (2013), Gehrer *et al.* (2012) and Neumann *et al.* (2013). Grades range from '1, very good' to '6, insufficient'. We are aware of the problem of

multicollinearity between grades and competency scores. Nonetheless, in Germany correlations between both quantities are comparably low in lower secondary school (e.g. around −0.32 between grades in German and reading competencies in 2012 (Authors Education Report (2016), page 94)). Thus, competencies are included in the outcome model to capture ability effects that grades do not map. As possible composition effect, we consider the proportion of mothers with a UAC in a school's entire ninth-grade level. The participation in the NEPS survey is voluntary. As a consequence, some schools in the sample have only a few ninth graders. We exclude schools with fewer than 10 ninth graders from our sample to reduce distortion and estimation problems. This results in a loss of 4% of the students. In sum, our data set comprises observations on 3291 ninth graders in 142 schools who were surveyed in 2011. The average number of ninth graders in a school is 23.2 (with a minimum of 10 students and a maximum of 48 students). In total, 77.2% of the ninth graders surveyed aspire to obtain a higher degree than they can obtain at *Hauptschule*. The grade level intraclass correlation ICC concerning higher aspirations of students is 22.7% (with a standard deviation of less than 0.1). Hence, the multilevel structure of our data is obvious. The on-line supplementary material shows all the model variables along with their mean values, standard deviations and the proportion of missing values. The variables on competencies and gender exhibit very few missing values (at maximum 4%), whereas the variables on migration background, aspirations and grades show a few more missing values (from 13% to 17%). However, we find a high percentage of missing values (more than 50%) for maternal education. Using Little's test (Little, 1988), we see that the missingness mechanism that generated our data set is not missingness completely at random, i.e. it is either MAR or MNAR. To cope with this issue, we use the FCS–MI approach, applying distinct imputation methods depending on the nature of the variable to be imputed. It is clear that a regressor in a multilevel model does not necessarily have to have a multilevel structure as well. Imputing missing values of a regressor with a single level structure by using a multilevel imputation model can lead to an unnecessarily high variance of the imputed values. Therefore, we first compute ICC for all the variables, using their observed values to assess whether a multilevel imputation routine is required. As a rule of thumb, we consider ICC-values that are higher than 20% as at risk of having a multilevel structure, whereas regressors with lower ICC-values are assigned a single-level imputation model. We find ICC-values larger than 20% for migration background and higher aspirations. Thus, higher aspirations and migration background are imputed by using multilevel imputation approaches, and all other variables are imputed by a single-level approach. From non-response analyses with similar NEPS data, we know that people with lower educational attainment are less likely to take part in the survey; see Zinn *et al.* (2018). Furthermore, we suspect that students with lower educational aspirations more often refuse to take part than their counterparts. Thus, we hypothesize that MNAR mechanisms generated the data on maternal education and educational aspirations. We use Galimard's imputation method (Galimard *et al.*, 2016) to impute the variable 'mother has UAC' and our novel method to impute 'higher aspirations'. As an exclusion criterion (in the selection models), we use the information on whether students were ever surveyed individually at home, on line or by phone—i.e. not at school—within nine waves (i.e. within 5 years). This appears to be an optimal choice since we find high correlations between this survey mode variable and the indicators of whether an aspiration value has been observed (around 95.8%) or of whether maternal education has been observed (around 46.1%). In contrast, we find low correlations between the survey mode variable and the observed aspiration values (0.03%) and the observed maternal education values (16.7%). Thus, we do not expect that the complete aspiration variable or the complete maternal education variable are substantially correlated with the survey mode variable. All other variables are imputed by using an MAR approach. In detail, missing values for grades are imputed by

Table 3. Effects on higher aspirations: analyses using different methods for handling missing values for maternal education

Variable	*Results for CCA*		*Results for MAR*		*Results for MNAR*	
	$\hat{\beta}$	*p-value*	$\hat{\beta}$	*p-value*	$\hat{\beta}$	*p-value*
Grade in mathematics	−0.112	0.203	−0.151	0.023	−0.151	0.018
Grade in German	−0.590	< 0.001	−0.472	< 0.001	−0.472	< 0.001
Competency in mathematics, satisfactory (reference category, poor)†	0.821	0.004	0.625	0.225	0.686	0.252
Competency in mathematics, good (reference category, poor)†	1.347	< 0.001	1.160	0.033	1.231	0.046
Competency in reading, satisfactory (reference category, poor)†	1.287	0.095	0.531	0.682	0.612	0.742
Competency in reading, good (reference category, poor)†	1.832	0.019	1.084	0.384	1.147	0.523
Sex (reference category, male)	0.843	< 0.001	0.748	< 0.001	0.756	< 0.001
Migration background (reference category, no)	0.832	< 0.001	0.549	< 0.001	0.490	< 0.001
Mother has UAC (reference category, no)	0.330	0.234	0.467	0.118	0.576	0.059
Proportion mothers with UAC in grade 9‡	2.756	0.003	3.885	0.011	3.927	0.013
Variance of random effect on grade level	0.474		0.948		0.949	
N students (in schools)	1250 (138)		3291 (142)		3291 (142)	

†Categories are created based on sample quantiles.
‡To compute this proportion only observed cases are used.

using predictive mean matching, and missing values for competence categories are imputed by using a polytomous regression approach. Missing values for gender are imputed by using a single-level logistic regression model, whereas missing values for migration background are imputed by using a two-level logistic regression model. Table 3 shows the results of our analysis, contrasted with the results of a CCA which is valid if missingness does not depend on observed or missing outcomes values given all other observed data (e.g. White and Carlin (2010)), and an MAR imputation approach for 'higher aspirations'. As in the simulation study, the two-level MAR imputation method is the method that was used in Zinn (2013). The number of imputed data sets is 20 with 50 iterations per imputed data set.

Under all three missing data schemes, we find significant effects (i.e. with $p < 0.05$) for higher grades in German, higher mathematics competencies, gender, migration background and the proportion of mothers with a UAC in the ninth grade. As expected, students with better grades in German show higher aspirations than students with lower grades. We do not find any significant effect of mathematics grades under a CCA. However, the effect of the mathematics grades becomes significant under MAR or MNAR. Thus, there is evidence that grades in mathematics are important for a student's aspirations. Having a look at the effect of competencies on students' aspirations, we see that the significance of German competencies shrinks to insignificance under MAR and MNAR—whereas, under a CCA, significant effects are estimated. For mathematics competencies, the picture is slightly different: even under MAR and MNAR, the significance of the good competency effect remains. In other words, there seems to be a significant part of remaining explanatory power in a student's mathematics competency that is not already captured by the mathematics grade (or vice versa). Apart from that, under all three missing data schemes, we find that female students in lower secondary school have much higher educational aspirations than males. Likewise, students with a migration background have much higher educational aspirations than those without. The proportion of ninth graders' mothers with a

UAC has a very strong effect on students' aspirations as well: the larger the proportion, the higher the effect. In sum, the size of this composition effect is very large. The reason for this lies in the small variance in the proportion of mothers with a UAC proportion at the grade level. It is less than 0.1. This small value is due to the target population under consideration, and not to particularities of the data. In Germany, social segregation in the lowest level of secondary education is enormously high. This is reflected in the data: in only one school is the proportion of ninth graders' mothers with a UAC higher than 50%, whereas, in 34.5% of the schools, none of the students has a mother with a UAC. Considering the effect of maternal education, we see that no significant influence can be detected under CCA and MAR. This changes under MNAR, although the effect of maternal education is statistically significant only at the 0.1-level. However, 58% of the cases with missing values concerning aspiration also have missing values for maternal education. Therefore, the chances are high that we are missing students with low aspirations and mothers without a UAC. Ignoring this issue may induce confounding bias, resulting in an underestimation of the effect of maternal education. In this respect, it is very likely that students who have a mother with a UAC have considerably higher educational aspirations than students whose mothers do not hold such a degree. To summarize, our analysis underscores the plausibility of the MNAR assumption concerning the 'maternal education' data. Having a look at the random-effect variances estimated under MAR and MNAR, the strong multilevel structure of the data becomes apparent. Neglecting this data feature when imputing missing values yields (at least) biased and misleading variance estimates. Exploring the missing data models that were estimated for maternal education and students' aspirations, we find a correlation (averaged across all imputations) ρ (between the error terms of the selection equation and the outcome model) of 0.25 for mother's education and of 0.06 for student's aspirations. This substantiates our prior finding of an MNAR mechanism having generated the 'maternal education' data (under the distributional assumptions of the selection model applied). In contrast, $\rho = 0.06$ indicates that an MAR mechanism generated the aspiration data (and not an MNAR mechanism). Looking additionally at the generally small differences between the estimates of the MAR and the MNAR models, this suspicion seems to be confirmed. The correlation τ between the random effects of the selection equation and the outcome model is on average 0.57. This high value suggests that participating in the study and having higher aspirations depend in a similar fashion on the (latent) class context of a student.

In situations where a single estimate is required (and not a set of estimates found within a sensitivity analysis), the results from the distinct imputation models of the sensitivity analysis can be combined by using Rubin's combining rules. This yields a multiple-model imputation approach that takes into account all sources of uncertainty with regard to the missing data mechanism that is assumed (Rubin, 1987; Siddique *et al.*, 2012, 2014). In our application, imputations are generated from different posterior predictive distributions (since we are using distinct (univariate) imputation models). This constitutes an additional source of uncertainty that must be considered in the final inference. Siddique *et al.* (2012) described a modified version of Rubin's combining rules based on nested MI (Rubin, 2003; Shen, 2000) that can be used for this.

5. Conclusion

In this paper, we introduced a novel and unique method for handling incomplete binary multilevel data that are assumed to be MNAR. Our univariate imputation method can easily be incorporated into the FCS framework to deal with multivariate missingness. For example, it

can be used in the R software package `mice`. (A related implementation is available from `http://github.com/AngelinaHammon/PaperBinaryMNARmultilevelData`.) Our simulation studies show that the novel approach outperforms competing techniques in terms of bias and coverage when data are affected by distinct MNAR mechanisms, i.e. in contrast with methods that are designed for MAR data, our method is capable of producing unbiased and accurate estimates of the quantities of interest. Moreover, our novel imputation method also yields valid estimates if the missing data were produced by an MAR mechanism. All in all, the method seems well suited to conducting meaningful sensitivity analyses—the only means of assessing the plausibility of an MNAR missing data mechanism or, more precisely, whether it matters for inference. In addition, the simulation studies demonstrate that valid imputation requires multilevel modelling if the data are clustered. In other words, when working with multilevel data that are supposed to be missing not at random, a proper imputation model like ours is required. In our simulation study, we kept the number of clusters and cluster sizes constant. A variation of both quantities can have an influence on the performance of our method. Thus, one of our future tasks will be to find out whether and to what extent such an influence exists. Furthermore, we included only random intercepts in our imputation model. The generalization to models containing random slopes is less straightforward and there is not a completely satisfactory imputation approach at the moment (Enders *et al.*, 2016; Grund *et al.*, 2016). However, in most social science applications, it is sufficient to consider only random intercepts to reflect a hierarchical structure in the data. We proved that our method is applicable to real data problems as well. For this, we studied the effect of maternal education on the educational aspirations of students in lower secondary education. However, we also noted that analysing large data sets with many clusters and incomplete predictors means long computing times when using only one processor. Thus, to run our approach, we highly recommend executing the MIs of `mice` in parallel on multiple cores.

Our approach must be extended in several respects. Up to now, we have applied a normal approximation for the parameter draws in our imputation algorithm. Using Bayesian estimation would make it possible to overcome this restriction since the model parameters can be drawn from the actual posterior distributions (which do not necessarily have to be normal). Furthermore, prior information on the correlation between the selection and the outcome model could be integrated. A decisive extension of the approach is the possibility of handling not only binary data but also, for example, ordinal variables and count data. The extension to ordinal data is straightforward since an ordinal outcome can also be explained by an underlying latent continuous variable. Therefore, an ordered probit model with sample selection (e.g. Greene (2012)) is a natural choice for imputing ordinal MNAR data. To perform meaningful sensitivity analyses, it is indispensable to compare alternative MNAR models with different assumptions regarding their missing data mechanisms. Thus, as a complement to a selection-model-driven approach, an imputation approach based on pattern–mixture modelling should be developed as well. For this, we plan to use the proxy pattern–mixture approach of Andridge and Little (2009, 2011). This method includes one sensitivity parameter to assess the robustness of the missing data inference. The parameter ranges between 0 and ∞, where 0 indicates MAR and ∞ means that missingness depends only on the incomplete variable. Since the sensitivity parameter is independent of any expert assessment, the method is well suited to social science applications. Finally, we point out the crucial role that is played by the exclusion criterion in the successful application of our method. The identification of a suitable variable for this task may seem difficult at first glance. Normally, however, when working with survey data, metainformation such as the survey mode or access corridors exists, which is suspected to be strongly correlated with the respondents' willingness to provide information, but not with the outcome variable to be

imputed. Nevertheless, it is advisable to carry out sensitivity analyses with regard to the exclusion criterion as well.

Acknowledgements

This paper uses data from the NEPS: starting cohort 4, DOI:10.5157/NEPS:SC4:9.1.1. From 2008 to 2013, NEPS data were collected as part of the 'Framework program for the promotion of empirical educational research' funded by the German Federal Ministry of Education and Research. From 2014, the NEPS has been carried out by the Leibniz Institute for Educational Trajectories at the University of Bamberg in co-operation with a nationwide network.

References

Abramowitz, M. and Stegun, I. A. (1964) *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. New York: Dover Publications.

Andridge, R. R. and Little, R. J. (2009) Extensions of proxy pattern-mixture analysis for survey nonresponse. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 2468–2482.

Andridge, R. R. and Little, R. J. (2011) Proxy pattern-mixture analysis for survey nonresponse. *J. Off. Statist.*, **27**, 153–180.

Asparouhov, T. (2006) General multi-level modeling with sampling weights. *Communs Statist. Theory Meth.*, **35**, 439–460.

Asparouhov, T. and Muthen, B. (2006) Multilevel modeling of complex survey data. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*

Audigier, V., White, I. R., Jolani, S., Debray, T., Quartagno, M., Carpenter, J., Van Buuren, S. and Resche-Rigon, M. (2017) Multiple imputation for multilevel data with continuous and binary variables. *Statist. Sci.*, **33**, 160–183.

Authors Education Report (2016) *Bildung in Deutschland 2016 [Education in Germany 2016]*. Bielefeld: Bertelsmann.

Bates, D., Mächler, M., Bolker, B. and Walker, S. (2015) Fitting linear mixed-effects models using lme4. *J. Statist. Softwr.*, **67**, 1–48.

Butler, J. S. and Moffitt, R. (1982) A computationally efficient quadrature procedure for the one-factor multinomial probit model. *Econometrica*, **50**, 761–764.

Cappellari, L. and Jenkins, S. P. (2003) Multivariate probit regression using simulated maximum likelihood. *Stata J.*, **3**, 278–294.

Davis, P. J. and Rabinowitz, P. (1967) *Numerical Integration*. Waltham: Blaisdell.

Delattre, E. and Moussa, R. (2015) On the estimation of causality in a bivariate dynamic probit model on panel data with Stata software: a technical review. *Working Paper 2015-04*. Université de Cergy-Pontoise, Cergy-Pontoise.

Duchhardt, C. and Gerdes, A. (2013) Scaling results of Starting Cohort 4 in ninth grade *Working Paper 22*. University of Bamberg, Bamberg.

Enders, C. K., Keller, B. T. and Levy, R. (2017) A fully conditional specification approach to multilevel imputation of categorical and continuous variables. *Psychol. Meth.*, **23**, 298–317.

Enders, C. K., Mistler, S. A. and Keller, B. T. (2016) Multilevel multiple imputation: a review and evaluation of joint modeling and chained equations imputation. *Psychol. Meth.*, **21**, 222–240.

Galimard, J.-E., Chevret, S., Curis, E. and Resche-Rigon, M. (2018) Heckman imputation models for binary or continuous MNAR outcomes and MAR predictors. *BMC Med. Res. Methodol.*, **18**, no. 1, article 90.

Galimard, J.-E., Chevret, S., Protopopescu, C. and Resche-Rigon, M. (2015) Imputation of MNAR missing data using one-step ML selection model. *36th A. Conf. International Society for Clinical Biostatistics*.

Galimard, J.-E., Chevret, S., Protopopescu, C. and Resche-Rigon, M. (2016) A multiple imputation approach for MNAR mechanisms compatible with Heckman's model. *Statist. Med.*, **35**, 2907–2920.

Gehrer, K., Zimmermann, C. and Weinert, S. (2012) The assessment of reading competence (including sample items for grade 5 and 9). *Research Data*. University of Bamberg, Bamberg.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. and Rubin, D. (2013) *Bayesian Data Analysis*. Boca Raton: Chapman and Hall–CRC.

Glynn, R., Laird, N. and Rubin, D. B. (1986) Selection modelling versus mixture modelling with nonignorable nonresponse. In *Drawing Inferences from Self-selected Samples* (ed. H. Wainer), pp. 115–142. New York: Springer.

Goldfarb, D. (1970) A family of variable-metric methods derived by variational means. *Math. Computn*, **24**, 23–26.

Golub, G. H. and Welsch, J. H. (1969) Calculation of Gauss quadrature rules. *Math. Computn*, **23**, 221–230.

Greene, W. H. (2004) Convenient estimators for the panel probit model: further results. *Empir. Econ.*, **29**, 21–47.
Greene, W. H. (2012) *Econometric Analysis*. Harlow: Pearson.
Grund, S., Lüdtke, O. and Robitzsch, A. (2016) Multiple imputation of missing covariate values in multilevel models with random slopes: a cautionary note. *Behav. Res. Meth.*, **48**, 640–649.
Hedeker, D., Mermelstein, R. J. and Demirtas, H. (2007) Analysis of binary outcomes with missing data: missing = smoking, last observation carried forward, and a little multiple imputation. *Addiction*, **102**, 1564–1573.
Kish, L. and Frankel, M. R. (1974) Inference from complex samples (with discussion). *J. R. Statist. Soc.* B, **36**, 1–37.
Larsen, R. (2011) Missing data imputation versus full information maximum likelihood with second-level dependencies. *Struct. Equn Modlng*, **18**, 649–662.
Lesaffre, E. and Spiessens, B. (2001) On the effect of the number of quadrature points in a logistic random-effects model: an example. *Appl. Statist.*, **50**, 325–335.
Little, R. (1988) A test of missing completely at random for multivariate data with missing values. *J. Am. Statist. Ass.*, **83**, 1198–1202.
Little, R. (1993) Pattern-mixture models for multivariate incomplete data. *J. Am. Statist. Ass.*, **88**, 125–134.
Little, R. (2008) Selection and pattern-mixture models. In *Longitudinal Data Analysis* (eds G. Fitzmaurice, M. Davidian, G. Verbeke and G. Molenberghs), pp. 409–431. Boca Raton: Chapman and Hall–CRC.
Liu, Q. and Donald, A. P. (1994) A note on Gauss-Hermite quadrature. *Biometrika*, **81**, 624–629.
Lüdtke, O., Robitzsch, A. and Grund, S. (2017) Multiple imputation of missing data in multilevel designs: a comparison of different strategies. *Psychol. Meth.*, **22**, 141–165.
Molenberghs, G. and Fitzmaurice, G. (2008) In *Longitudinal Data Analysis* (eds G. Fitzmaurice, M. Davidian, G. Verbeke and G. Molenberghs), pp. 395–408. Boca Raton: Chapman and Hall–CRC.
Molenberghs, G., Verbeke, G. and Kenward, M. (2008) In *Longitudinal Data Analysis* (eds G. Fitzmaurice, M. Davidian, G. Verbeke and G. Molenberghs), pp. 501–551. Boca Raton: Chapman and Hall–CRC.
Morris, T. P., White, I. R. and Crowther, M. J. (2019) Using simulation studies to evaluate statistical methods. *Statist. Med.*, **38**, 2074–2102.
Mulkay, B. (2015) Bivariate probit estimation for panel data: a two-step Gauss-Hermite quadrature approach with an application to product and process innovations for France. Université de Montpellier, Montpellier.
Naylor, J. C. and Smith, A. F. M. (1982) Applications of a method for the efficient computation of posterior distributions. *Appl. Statist.*, **31**, 214–225.
Neumann, I., Duchhardt, C., Grüäŷing, M., Heinze, A., Knopp, E. and Ehmke, T. (2013) Modeling and assessing mathematical competence over the lifespan. *J. Educ. Res. Online*, **5**, 80–109.
Nocedal, J. and Wright, S. (2006) *Numerical Optimization*. New York: Springer Science and Business Media.
Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H. and Rasbash, J. (1998) Weighting for unequal selection probabilities in multilevel models. *J. R. Statist. Soc.* B, **60**, 23–40.
Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J. and Solenberger, P. (2001) A multivariate technique for multiply imputing missing values using a sequence of regression models. *Surv. Methodol.*, **27**, 85–96.
Rendtel, U. (1992) On the choice of a selection-model when estimating regression-models with selectivity. *Discussion Paper*. German Institute for Economic Research, DIW, Berlin.
Resche-Rigon, M. and White, I. R. (2018) Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Statist. Meth. Med. Res.*, **27**, 1634–1649.
Resseguier, N., Giorgi, R. and Paoletti, X. (2011) Sensitivity analysis when data are missing not-at-random. *Epidemiology*, **22**, 282.
Rubin, D. B. (1974) Characterizing the estimation of parameters in incomplete-data problems. *J. Am. Statist. Ass.*, **69**, 467–474.
Rubin, D. B. (1976) Inference and missing data. *Biometrika*, **63**, 581–592.
Rubin, D. B. (1977) Formalizing subjective notions about the effect of nonrespondents in sample surveys. *J. Am. Statist. Ass.*, **72**, 538–543.
Rubin, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
Rubin, D. B. (2003) Nested multiple imputation of NMES via partially incompatible mcmc. *Statist. Neerland.*, **57**, 3–18.
Schneider, E. (2018) *Von der Hauptschule in die Sekundarstufe II: eine Schülerbiografische Längsschnittstudie*. Berlin: Springer.
Shen, Z. (2000) Nested multiple imputations. *PhD Thesis*. Harvard University, Cambridge.
Siddique, J., Harel, O. and Crespi, C. M. (2012) Addressing missing data mechanism uncertainty using multiple-model multiple imputation: application to a longitudinal clinical trial. *Ann. Appl. Statist.*, **6**, 1814–1837.
Siddique, J., Harel, O., Crespi, C. M. and Hedeker, D. (2014) Binary variable multiple-model multiple imputation to address missing data mechanism uncertainty: application to a smoking cessation trial. *Statist. Med.*, **33**, 3013–3028.
Tompsett, D. M., Leacy, F., Moreno-Betancur, M., Heron, J. and White, I. R. (2018) On the use of the not-at-random fully conditional specification (NARFCS) procedure in practice. *Statist. Med.*, **37**, 2338–2353.
Train, K. E. (2009) *Discrete Choice Methods with Simulation*. New York: Cambridge University Press.
Van Buuren, S., Boshuizen, H. C. and Knook, D. L. (1999) Multiple imputation of missing blood pressure covariates in survival analysis. *Statist. Med.*, **18**, 681–694.

Van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C. G. M. and Rubin, D. B. (2006) Fully conditional specification in multivariate imputation. *J. Statist. Computn Simuln*, **76**, 1049–1064.
Van Buuren, S. and Groothuis-Oudshoorn, K. (2011) mice: multivariate imputation by chained equations in R. *J. Statist. Softwr.*, **45**, 1–67.
Von Hippel, P. T. (2007) Regression with missing *y*s: an improved strategy for analyzing multiply imputed data. *Sociol. Methodol.*, **37**, 83–117.
Warm, T. A. (1989) Weighted likelihood estimation of ability in item response theory. *Psychometrika*, **54**, 427–450.
White, I. R. and Carlin, J. B. (2010) Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statist. Med.*, **29**, 2920–2931.
Wooldridge, J. (2002) *Econometric Analysis of Cross Section and Panel Data*. Cambridge: MIT Press.
Wößmann, L. (2007) Fundamental determinants of school efficiency and equity: German states as a microcosm for OECD countries. *Discussion Paper 2880*. Institute of Labor Economics, Bonn.
Wothke, W. (2000) In *Modeling Longitudinal and Multilevel Data: Practical Issues, Applied Approaches, and Specific Examples* (eds T. Little, K. Schnabell and J. Baumert), pp. 219–240, 269–281. Mahwah: Erlbaum.
Zinn, S. (2013) An imputation model for multilevel binary data. *Working Paper 31*. National Educational Panel Study, University of Bamberg, Bamberg.
Zinn, S., Würbach, A., Steinhauer, H. and Hammon, A. (2018) Attrition and selectivity of the NEPS starting cohorts: an overview of the past 8 years. *Survey Paper 34*. Leibniz Institute for Educational Trajectories, National Educational Panel Study, Bamberg.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article.

'Supplementary material to "Multiple imputation of binary multilevel missing not at random data"'.