A comparison of incomplete-data methods for categorical data

van der Palm, D.W.; van der Ark, L.A.; Vermunt, J.K.

Published in:
Statistical Methods in Medical Research

DOI:
10.1177/0962280212465502

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (http://dare.uva.nl)
A comparison of incomplete-data methods for categorical data

Daniël W van der Palm, L Andries van der Ark and Jeroen K Vermunt

Abstract
We studied four methods for handling incomplete categorical data in statistical modeling: (1) maximum likelihood estimation of the statistical model with incomplete data, (2) multiple imputation using a loglinear model, (3) multiple imputation using a latent class model, (4) and multivariate imputation by chained equations. Each method has advantages and disadvantages, and it is unknown which method should be recommended to practitioners. We reviewed the merits of each method and investigated their effect on the bias and stability of parameter estimates and bias of the standard errors. We found that multiple imputation using a latent class model with many latent classes was the most promising method for handling incomplete categorical data, especially when the number of variables used in the imputation model is large.

Keywords
Missing data, categorical data, multiple imputation, latent class analysis, MICE, maximum likelihood, medical research

1 Introduction
This paper discusses methods to handle incomplete categorical data. Many medical studies deal solely with analyzing categorical data and, consequently, the statistical model that is used to analyze the data (from here on referred to as the substantive model) is also tailored to categorical data. For example, predictors of reduced length of hospital stay were studied using logistic regression, determinants of caregivers’ health were studied using loglinear modeling, and the effectiveness of the World Health Organization Disability Assessment Schedule II was investigated using a nonparametric item response analysis. A frequently encountered problem is that the data are incomplete, which prevents a straightforward statistical analysis; a researcher should handle this problem appropriately. Klebanoff and Cole found that the majority of applied researchers resort to ad-hoc methods such as complete-case analysis or pair-wise deletion, which may lead to biased statistical results and reduced power.5,6

Corresponding author:
Daniël W van der Palm, Tilburg School of Social and Behavioral Sciences, Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands.
Email: d.w.vd.palm@uvt.nl
For handling incomplete continuous data, adequate alternatives have been proposed, extensively researched, and implemented in major software packages such as SPSS and SAS. Hence, there is no need for applied researchers to resort to ad-hoc methods in case of continuous data.

Incomplete data methods for categorical data have not yet been crystallized out, and it is unknown which method should be recommended to practitioners. Ideally, an incomplete-data method should meet three criteria. For the substantive model, it should produce parameter estimates (i) that are unbiased, (ii) that are stable in order to avoid unnecessary loss of power in the statistical analysis, and (iii) that have standard errors correctly reflecting the uncertainty due to missing data. Ideally, these criteria should be met for data sets with both small and large numbers of variables, sample sizes, and percentages of incomplete data, and for both simple and complex associations in the data.

With respect to these criteria, two incomplete-data methods for categorical data are especially promising: Multiple imputation using latent class analysis (MILC) and multivariate imputation using chained equations (MICE). Both methods have the practical advantage that they can easily handle data sets containing a large number of variables and respondents. However, researchers having incomplete categorical data cannot yet readily apply MILC and MICE because there are various unresolved issues (explained hereunder). The impact of these issues on the three criteria for substantive models is unknown. In this study, we discuss two reasonable options for the unresolved issues for both MILC and MICE, and investigate to which degree they meet the three criteria, so as to decide which incomplete-data method should be selected for categorical data. Multiple imputation using a loglinear model (MILL) and maximum likelihood for incomplete data (MLID) are used as benchmarks. MILL is known to produce unbiased parameter estimates but can only handle a small number of variables; MLID is known to be asymptotically unbiased but may run into difficulties as the number of variables becomes very large.

The remainder of this paper is organized as follows. First, we briefly discuss the four incomplete-data methods. For both MILC and MICE, we discuss two variants, resulting in six incomplete-data methods in total. Second, we compare the advantages and disadvantages of the methods in a theoretical discussion. Third, we present the results of two simulation studies. In Study 1, for dichotomous data, we compared MILC, MICE, MILL, and MLID with respect to the three criteria. In Study 2, for trichotomous data, we compared MILC, MICE, and complete-case analysis with respect to the three criteria. Fourth, we applied MLID, MILC, MICE, and complete-case analysis to a medical data set. Finally, we give recommendations based on the theoretical discussion and the two simulation studies.

2 Incomplete-data methods

2.1 Incomplete data

Let \( Y = (Y_1, Y_2, \ldots, Y_J) \) denote the scores on the \( J \) variables, and let \( \theta \) be the generic notation for the vector of unknown parameters of the joint distribution of \( Y \), denoted \( P(Y; \theta) \). To distinguish specific models Greek letters other than \( \theta \) may also be used to denote parameter vectors. Note that \( Y_j \) may be either a predictor variable or an outcome variable depending on the substantive model. If confusion arises, we add the superscripts \( p \) and \( o \) to indicate that a variable serves as a predictor variable or outcome variable, respectively. \( Y \) may contain missing values, and the objective is to deal with them appropriately.

Most incomplete-data methods, including the ones considered in this paper, assume that the mechanism that caused the missing values is ignorable, which means that two conditions should
hold. First, the parameters that govern the missing data process must be unrelated to the parameters to be estimated, which is a rather unrestrictive assumption. Second, the data must be missing at random (MAR), which means that whether or not a score is missing only depends on scores observed in the study. If, after conditioning on all observed data, the missingness depends on missing values of variables included in the study or on variables not included in the study, MAR is violated and, as a result, the missingness mechanism is non-ignorable. Non-ignorable missingness may cause biased parameters in the substantive model (first criterion). Apart from special studies with planned missingness, MAR is unlikely to hold in practice, and it is impossible to test whether the MAR assumption holds for a particular data set. Therefore, the degree to which MAR is violated (i.e. the degree to which the observed scores cannot explain the missingness mechanism) becomes important: If the violation of MAR becomes more severe, the parameter bias in the substantive model is likely to increase. If the number of variables in a data set increases, the degree to which the variables can explain the missingness mechanism is also likely to increase. Hence, if an incomplete-data method can handle a large number of variables, and if a large number of variables is available, the violation of MAR will most likely be less severe and the missingness mechanism is more likely to be ignorable. This notion plays an important role in our evaluation of incomplete-data methods and will be referred to as Schafer’s notion on the number of variables.

2.2 Description of incomplete-data methods

2.2.1 Maximum likelihood for incomplete data
MLID is a well-known and documented method to obtain parameter estimates and standard errors in the presence of missing data. MLID constitutes estimating the parameters of the substantive model and their standard errors, using all observed data. For example, when studying predictors of reduced length of hospital stay using logistic regression, MLID can be used to estimate the logistic regression model using all observed data. No further action is required; the obtained parameter estimates and standard errors can be directly interpreted. The substantive model can be an asymmetric model such as a logistic regression model or an item response theory model, which describe the conditional distribution of the outcome variables given the predictor variables $P(Y_o|Y_p; \theta)$, or a symmetric model, such as a loglinear model, latent class model, or canonical correlation model, which describe the joint distribution of all variables $P(Y; \theta)$. MLID assumes that the missingness mechanism is ignorable. For categorical data, specialized software is usually required to conduct MLID, such as LEM or Mplus.

2.2.2 Multiple imputation
Multiple imputation consists of creating $m$ completed data sets by replacing the missing values in the data with plausible values $m$ times. These plausible values replacing the missing values are called the imputed values. The statistical model that generates imputed values is referred to as the imputation model. After the multiple imputation, on each of the $m$ completed data sets a substantive model is estimated, and the $m$ sets of parameter estimates and standard errors are combined into a single set. Most researchers use $m = 5$, but this value is currently debated. Using multiple imputation allows for separating the missing data handling and the substantive analysis; a researcher can estimate substantive models as if there had been no missing data, or distribute the completed data to other researchers for further analysis.

Multiple imputation starts in the same way as MLID for symmetric models: A statistical model is estimated describing the joint distribution $P(Y; \theta)$. Rather than a substantive model, this model is an imputation model for obtaining imputed values from $P(Y; \theta)$. For example, when studying
predictors of reduced length of hospital stay using logistic regression, a loglinear model describing the joint distribution of both predictor variables and reduced length of hospital stay may be used as an imputation model to generate imputed values replacing the missing data \( m \) times. After the multiple imputation, logistic regression analysis can be conducted on the completed data sets.

One must account for the fact that the imputed values are not observed and, therefore, uncertain. There are two sources of uncertainty. Firstly, the estimated parameters of the imputation model are uncertain; this uncertainty is expressed by their standard errors. Secondly, there is uncertainty due to sampling variability when drawing imputed values from \( P(Y; \theta) \). To account for parameter uncertainty, for each of the \( m \) data sets, a different set of parameters of the imputation model is used. In a Bayesian framework, the \( m \) sets of parameters of the imputation models are random draws from \( P(\theta|Y) \), the distribution of the parameters given the data. In a frequentist framework, the \( m \) sets of parameters are estimated using \( m \) nonparametric bootstrap samples of the data. A nonparametric bootstrap sample consists of randomly drawing a new sample of \( N \) observations with replacement. To reflect uncertainty due to sampling variability, the replacement of missing values is done \( m \) times, yielding \( m \) completed data sets. The three multiple imputation methods for categorical data discussed in this paper differ in the way that they describe the joint distribution, \( P(Y; \theta) \), and how they account for parameter uncertainty. MILL is discussed briefly because this method is ready for use; MILC and MICE are described in more detail so as to allow the discussion of the specific options these methods offer.

**Multiple imputation using a loglinear model.** MILL uses a loglinear model as the imputation model. Let the parameters of the loglinear model be denoted \( \lambda \); the saturated loglinear model for dichotomous responses can be written as

\[
\log P(Y; \lambda) = \lambda + \sum_{i=1}^{J} \lambda_i Y_i + \sum_{i=1}^{J-1} \sum_{j=i+1}^{J} \lambda_{ij} Y_i Y_j + \cdots + \lambda_{1,2,\ldots,J} Y_1 Y_2 \cdots Y_J. \tag{1}
\]

The joint distribution is obtained by taking the exponential of the right-hand side of equation (1). Typically, a saturated loglinear model is used to obtain imputation values because it captures all possible associations in the data; therefore, it is the gold standard for multiple imputation of categorical data. If higher-order interaction terms are omitted, the approximation of the joint distribution by the loglinear model may deteriorate. MILL can, for example, be conducted using software packages CAT or Latent GOLD which utilize a Bayesian and a nonparametric bootstrap approach, respectively, to account for parameter uncertainty.

**Multiple imputation using a latent class model.** MILC uses a latent class model to estimate the joint distribution of the variables in the data. Let \( X \) denote a discrete latent variable with \( K \) latent classes, indexed by \( k (k = 1, \ldots, K) \). Let \( \pi \) denote the vector of parameters of the latent class model; \( \pi \) can be divided into \( \pi_x \), the latent class proportions, and \( \pi_y \) the conditional response probabilities. Under a latent class model, joint distribution \( P(Y; \pi) \) has the following form:

\[
P(Y; \pi) = \sum_{k=1}^{K} P(X = k; \pi_x) P(Y|X = k; \pi_y) = \sum_{k=1}^{K} P(X = k; \pi_x) \prod_{j=1}^{J} P(Y_j|X = k; \pi_y).
\]
If the number of latent classes is sufficiently large, a latent class model correctly picks up the first, second, and higher order moments of the response variables, as is the case with all forms of mixture models. It is unknown how many latent classes are sufficient for a good approximation of the joint distribution. Vermunt et al. argued that it is better to have too many than too few latent classes. Therefore, out of three selection criteria, Akaike’s information criterion (AIC), Bayesian information criterion (BIC), and AIC3, they suggested using AIC to select the number of latent classes because it yields the largest number of latent classes. Hence, letting AIC determine the number of latent classes, abbreviated MILC (AIC), is the first option for MILC. However, it is expected that an even larger number of latent classes can further improve the approximation of the joint distribution. Having a relatively large number of latent classes, abbreviated MILC (Large), is the second option for MILC. MILC can be applied using Latent GOLD, which uses the nonparametric bootstrap to account for parameter uncertainty.

**Multivariate imputation using chained equations.** MICE is a fully conditional specification method, which specifies the imputation model on a variable-by-variable basis using a separate conditional distribution for each incomplete variable. Let $Y_{-j}$ denote the scores on all variables except $Y_j$. MICE reduces the problem of finding one $J$-dimensional joint distribution $P(Y; \theta)$ to finding $J$ univariate conditional distributions $P(Y_1|Y_{-1}; \theta), \ldots, P(Y_J|Y_{-J}; \theta)$. Conditional distribution $P(Y_j|Y_{-j}; \theta)$ is used for imputation of $Y_j$ ($j = 1, \ldots, J$). Under certain conditions, a draw from each of the $J$ conditional distributions is equivalent to a single draw from the joint distribution, but it is not guaranteed. Results from simulation studies suggest that the problem is unlikely to be serious in practice. MICE starts with replacing missing values of the variables by draws from their respective marginal distributions. Next, in an iterative process, the imputed values are updated variable by variable using the univariate conditional distributions. When $Y_j$ is imputed, the other variables act as predictor. If the joint distribution that is defined by the $J$ conditional distributions exists then this iterative process is a Gibbs sampler and converges to the joint distribution of the $J$ variables. Often, as little as 10 to 20 iterations are required.

The imputation model describing the conditional probabilities $P(Y_1|Y_{-1}; \theta), \ldots, P(Y_J|Y_{-J}; \theta)$ can be any appropriate regression model depending on the nature of the outcome variable: linear regression in combination with predictive mean matching, logistic regression, polytomous regression, and nonlinear regression. We focused on two imputation models; the first one being logistic regression (abbreviated MICE (LOG)) which is the default method in the R-package MICE for dichotomous outcome variables (for Study 2, polytomous regression is used, which is the extension of MICE (LOG) to variables with more than 2 categories; for details see, e.g. Van Buuren et al.). Let $\beta$ denote the vector of parameters for the logistic regression model. MICE (LOG) models conditional distribution $P(Y_j|Y_{-j}; \beta)$ as

$$
\logit[P(Y_j|Y_{-j}; \beta)] = \beta_0 + \beta_1 Y_1 + \cdots + \beta_{j-1} Y_{j-1} + \beta_{j+1} Y_{j+1} + \cdots + \beta_J Y_J
$$

We also considered linear regression in combination with predictive mean matching (abbreviated MICE (PMM)). The first step of MICE (PMM) is to obtain a predicted value by means of linear regression in which all other variables serve as predictors. In the second step, the respondent that has the most similar predicted value as well as an observed value on the variable that is being imputed is selected as the nearest neighbor. Subsequently, the observed value of this nearest neighbor is used as the imputation value for the respondent with a missing value.
Parameter uncertainty is accounted for in a Bayesian framework; a new set of parameters is drawn from its posterior distribution for the construction of each imputed data set. More specifically, the MICE algorithm involves iteratively sampling parameter values $\beta$ from their posterior distribution and imputing the missing values $Y_j$ by drawing from the conditional distribution $P(Y_j|Y_{-j}; \beta)$. This corresponds with a Gibbs sampling scheme if the joint distribution of the variables can be constructed from their univariate conditional distributions and if the distribution from which parameters are drawn can be constructed from the joint distribution of the variables and an appropriate prior distribution. These two conditions are not fulfilled when using MICE with categorical data, which means that the algorithm is not an exact Gibbs sampler. MICE can be conducted using the R package MICE or the STATA package ICE.

2.2.3 Other incomplete-data methods

We have three remarks on other incomplete-data methods. First, besides MLID and multiple imputation, there are two other categories of incomplete-data methods: the fully Bayesian method and weighted estimating equations. We did not consider these two approaches to limit the scope of this paper. A full Bayesian analysis with for example WinBugs is in fact similar to both MLID and multiple imputation; that is, the parameters of the substantive model are estimated using the incomplete data using an algorithm containing a step in which the missing values are imputed. Results can be expected to be similar to MLID. Weighting is typically used to deal with completely missing data and has limited practical use with partially missing data. It may moreover yield unstable estimates in the presence of influential weights.

Second, a popular imputation model for multiple imputation is the multivariate normal distribution. The method is robust against deviations from normality and may even perform well for categorical data, although some studies reported serious bias. We did not consider incomplete data-methods that were not designed for categorical data as these methods are not suitable for nominal variables (e.g., blood type, eye color, surgical outcome).

Third, the best known ad-hoc method is probably complete-case analysis, in which only the observations without any missing values are used to estimate the substantive model. In other words, subjects who have at least one missing value are discarded from the analysis. Hence, in contrast to MLID, complete-case analysis does not incorporate all available information. Complete-case analysis reduces power and may yield biased parameter estimates for the substantive model if the data are not missing completely at random (MCAR); this MCAR assumption is considered to be unrealistic in most situations. Complete-case analysis is included in Study 2 and the real-data example. For Study 2, the number of variables was too large for more preferable benchmarks such as MILL and MLID.

2.3 Advantages, disadvantages, and unresolved issues of the incomplete-data methods

2.3.1 Practical issues

For application of the incomplete-data methods, sample size, complexity of the association structure in the data, and percentage of missingness are not restrictive for any of the methods. A limitation of MILL is that it cannot handle large numbers of variables because the number of cells in the contingency table that has to be evaluated in the loglinear model becomes too large. For example, the number of cells that need be evaluated exceeds one million for 20 dichotomous variables and one billion for 30 dichotomous variables, 19 trichotomous variables, or 13 variables with five categories. In cases where the substantive model contains fewer variables than available
in the data set, a possible solution is to consider only those variables that are used in the substantive model. However, following Schafer’s notion on the number of variables, using only a small number of variables for the imputation model may result in biased parameter estimates. For MILC and MICE, large numbers of variables do not pose a problem. A potential problem for MLID is that it usually requires specialized software, depending on the substantive model that one wants to estimate, whereas standard data-analysis techniques can be applied after the imputation phase of MILL, MILC and MICE. Moreover, MLID can only be used if the number of variables in substantive model is not too large.

2.3.2 Bias
We consider three possible causes of bias in the parameter estimates: First, non-ignorable missingness in the data. Following Schafer’s notion on the number of variables, it is suggested that the inclusion of many variables in the imputation model makes it more likely that violations of ignorability are minor. The second possible cause of bias is misspecification of the imputation model so that it is too parsimonious. The imputation model should be as general as possible; this ensures that the imputed values behave as neutral as possible in subsequent analyses. Hence, the main criterion of an adequate imputation model is whether it captures all the associations between categorical variables that exist on the population level. The third possible cause of bias is misspecification of the substantive model. However, this is unrelated to the incomplete-data method being used and is not pursued further.

For MLID, no imputation model needs to be specified but a violation of the ignorability assumption may result in biased parameter estimates. Statistical analyses that are based on MLID only include those variables in the data that are substantively relevant, possibly excluding many variables. When the number of variables in the substantive model is small, then, following Schafer’s notion on the number of variables, the missingness mechanism in the reduced data is less likely to be ignorable. Simulation studies showed that under ignorable missingness, MLID yields unbiased parameter estimates.

For MILL, the imputation model being too parsimonious is not an issue because the imputation model is typically the saturated model. However, MILL can handle only a limited number of categorical variables. As a result, following Schafer’s notation on the number of variables, the missingness mechanism in the reduced data may not be treated as ignorable possibly resulting in biased parameter estimates. Simulation studies showed that under ignorable missingness, MILL yields unbiased parameter estimates.

For MILC and MICE, the amount of non-ignorable missingness may be reduced if the data contain many variables relevant for predicting the missing values (Schafer’s notion on the number of variables) because both methods can handle a very large number of (auxiliary) variables. For MICE, it is unknown which of the two variants yields the least bias, for MILC, it is expected that a large number of latent classes, MILC (Large), produces less bias than a smaller number of latent classes, MILC (AIC).

2.3.3 Stability
We consider three possible causes that influence the stability of parameter estimates in the presence of incomplete data. A first possible cause is a too small effective sample. It is well known sample size has a positive effect on stability. None of the incomplete-data methods under investigation unduly reduce the effective sample size, in the way some ad hoc methods do (e.g., complete-case analysis, pair-wise deletion). However, it is unknown whether the incomplete-data methods under investigation yield the same stability of parameter estimates given a fixed sample size. A second
possible cause is misspecification of the imputation model so that it is too complex. This is the well-
known tradeoff between bias and stability: If the imputation model is too parsimonious it may result
in biased outcomes, if it is too complex, it may result in less stable outcomes. For most researchers,
unbiased parameter estimates are more important than stable parameter estimates. The third
possible cause of instability is misspecification of the substantive model so that it is too complex.
However, this is unrelated to the incomplete-data method being used and is not pursued further.

Only for the second possible cause of instability, an overly complex imputation model, we have
expectations for the incomplete-data methods under investigation. MLID does not require an
imputation model, so no loss of stability can ensue from an overly complex imputation model.
For MILL, the imputation model is saturated meaning that it is expected to be overly complex in
most cases. Therefore, a certain loss of stability is expected for MILL in comparison to MLID.

For MILC, the two variants are expected to differ in stability because their respective imputation
models differ in complexity. MILC (Large) uses a relatively large number of latent classes which
means that its imputation model is expected to be able to capture every possible association. As is
the case with MILL, results produced by MILC (Large) are expected to lose a certain degree of
stability because its imputation model is expected to be overly complex. MILC (AIC) estimates the
required number of latent classes using AIC, which results in a relatively small number of latent
classes. Therefore, its imputation model is more parsimonious and its results are expected to be more
stable than MILC (Large).

For MICE, the two variants differ in the conditional imputation model that is used. The stability of
MICE depends on the degree to which higher order associations are included in the conditional
imputation model. The default setting of MICE (PMM) only includes main effects. However,
because predictive mean matching is used, all associations can be picked up for data sets with a small
number of variables. Therefore, we expect that the stability of the parameter estimates produced by
MICE (PMM) is similar to MILL and MILC (Large). The default setting of MICE (LOG) also only
includes main effects. Therefore, MICE (LOG) is expected to have relatively stable results.

2.3.4 Bias of the standard errors
It is unknown whether the six incomplete-data methods overestimate or underestimate the standard
errors of parameter estimates. Hence, we have no specific expectations with regard to the bias of the
standard errors.

3 Study 1: Bias, stability, and bias in standard errors produced by MILC,
MICE, MILL, and MLID for a small number of dichotomous variables

In Study 1, we compared incomplete-data methods MILC (AIC), MILC (Large), MICE (PMM),
and MICE (LOG) to MLID and MILL, on bias of the parameter estimates, stability of the
parameter estimates, and bias of the standard errors. Because MLID and MILL can handle only
a limited number of variables, the number of variables was kept small. The design of Study 1 was
motivated by the study of Kurian et al.,1 who studied several predictors of a single outcome variable
reduced “length of hospital stay” using logistic regression.

3.1 Method

3.1.1 General set up
The set up of the simulation study was as follows. First, we sampled complete data sets from a
population model. Second, we created incomplete data sets by deleting variable scores according to
an MAR missingness mechanism. Third, for each incomplete data set we constructed five completed data sets using a missing-data method. Fourth, we used the completed data sets to estimate the parameters of the regression part of the population model and we reported the bias and stability of the parameter estimates.

The population model was defined for five dichotomous predictor variables \( Y_1, \ldots, Y_5 \), and one dichotomous outcome variable \( Y_6 \). The categories were coded 0 and 1 (dummy coding). Dummy coding was used because it is the most commonly used coding scheme for logistic regression models. The associations among the predictor variables \( Y_1, \ldots, Y_5 \) were described by loglinear model

\[
\log P(Y_1, Y_2, Y_3, Y_4, Y_5) = -1.47 + \sum_{j=1}^{5} -2 \cdot Y_j + \sum_{j=1}^{4} \sum_{k=j+1}^{5} 1 \cdot Y_j Y_k
\]  

(2)

Outcome variable \( Y_6 \) was related to the predictor variables by logit model

\[
\logit(Y_6) = \beta_0 + \beta_1 Y_1 + \beta_2 Y_2 + \beta_3 Y_3 + \beta_4 Y_4 + \beta_5 Y_5 - \beta_{23} Y_2 \cdot Y_3,
\]  

(3)

which contains main effects of the predictor variables as well as the interaction effect of \( Y_2 \) and \( Y_3 \). The strength of the interaction term, \( \beta_{23} \), was manipulated in the study. The coefficients \( \beta_0, \beta_2, \text{and} \beta_3 \) are changed together with \( \beta_{23} \) so that the average logit and the average effects of \( Y_2 \) and \( Y_3 \) remain constant across conditions. Complete data sets were created by sampling from \( P(Y_1, Y_2, Y_3, Y_4, Y_5) \) (equation (2)) and \( P(Y_6|Y_1, Y_2, Y_3, Y_4, Y_5) \) (equation (3)).

Variables \( Y_1 \) and \( Y_2 \) had missing values that were MAR. Variables \( R_1 \) and \( R_2 \) indicated whether a score was missing, \( R_i = 0 \), or observed, \( R_i = 1 \), for \( Y_1 \) and \( Y_2 \), respectively. Missing values in \( Y_1 \) were created using logistic regression model

\[
\logit(R_1) = \gamma_1 + 1.09 \cdot Y_3 + 2.01 \cdot Y_4 - 0.79 \cdot Y_3 Y_4,
\]  

(4)

and missing values in \( Y_2 \) were created using logistic regression model

\[
\logit(R_2) = \gamma_2 + 1.04 \cdot Y_5 + 1.94 \cdot Y_6 - 0.74 \cdot Y_5 Y_6
\]  

(5)

The total percentage of missingness (one of the predictor variables in Study 1, to be discussed later) was manipulated by changing the intercepts (\( \gamma_1 \) and \( \gamma_2 \)) in equations (4) and (5). This approach allows for varying the total percentage of missingness without altering the strength of associations between the predictor variables and the missingness indicator variable in equations (4) and (5).

For each incomplete data set, five completed data sets were created using a multiple imputation method and for each completed data set logistic regression model

\[
\logit(Y_6) = \beta_0 + \beta_1 Y_1 + \beta_2 Y_2 + \beta_3 Y_3 + \beta_4 Y_4 + \beta_5 Y_5 + \beta_{23} Y_2 Y_3
\]  

(6)

was estimated. Rubin’s\textsuperscript{23} rules were used to combine the five sets of regression parameter estimates. It should be noted that \( m = 5 \) completed data sets is usually considered to be sufficient to obtain stable results.\textsuperscript{6} However, other researchers have argued that \( m \) should be based on the total percentage of missingness; for example, \( m \) should equal the total percentage of missing data to obtain a sufficient degree of stability in the results.\textsuperscript{22} In many cases, this would render \( m \) larger than five. We note that this is especially important for a single analysis; in a simulation study the size of \( m \) has much less influence because of the large number of replications.
Three software packages were used for multiple imputation and parameter estimation. Data were generated using software package LEM, methods MILC and MILL were conducted using the software program Latent GOLD 4.5, and for MICE we used the R package MICE. After multiple imputation, the substantive model, defined in equation (6), was estimated using Latent GOLD 4.5. For MLID, the substantive model was estimated using LEM.

3.1.2 Predictor variables
Incomplete-data method was a within factor with six levels: MILC (AIC), MILC (Large), MICE (PMM), MICE (LOG), MLID, and MILL. The incomplete-data methods determine the imputation model and may thus affect both bias and stability.

Strength of the interaction term was a between factor with three levels that was manipulated by varying parameter $\beta_{23}$ in equation (3). The levels were no three-way association ($\beta_{23} = .00$), medium ($\beta_{23} = -2.00$), and strong ($\beta_{23} = -2.00$). Strength of the three-way association sets requirements for the complexity of the imputation model. If this effect increases, a more complex imputation model is required to pick up the associations in the data. It is expected that strength of the three-variable association affects both bias and stability.

Percentage of missingness was a between factor with three levels: moderate (10% missingness), high (20% missingness), and extreme (40% missingness). The percentage of missingness was manipulated by varying parameters $\gamma_1$ and $\gamma_2$ in equations (4) and (5), respectively. For 10% missingness, $\gamma_1 = -2.46$ and $\gamma_2 = -2.53$, for 20% missingness, $\gamma_1 = -1.41$ and $\gamma_2 = -1.44$, and for 40% missingness, $\gamma_1 = -.39$ and $\gamma_2 = -.41$. As the percentage of missingness increases, the imputation model becomes more important. The condition with 40% of missingness is included because the consequences of an inadequate imputation model are magnified by an increase in the percentage of missingness.

Sample size was a between factor with two levels: Small ($N = 200$) and large ($N = 1000$). Sample size is expected to predominantly affect stability. In particular, the aim was to examine how sample size is related to the stability of the statistical results in the analysis of interest for each missing-data method.

The four predictor variables were fully crossed producing a $6 \times 3 \times 3 \times 2$ design, with 1000 replications for each of the 18 combinations of the between-subjects variables.

3.1.3 Outcome variables
The outcome variables were bias of parameter estimates, standard deviation of parameter estimates across replications, and bias of the reported standard errors. Let $\hat{\beta}_{bj}$ denote a parameter estimate of the $j$th variable (equation (6)) in replication $b$ ($b = 1, \ldots, 1000$), then the bias over 1000 replications was computed as

$$\text{bias} = \frac{1}{1000} \sum_{b=1}^{1000} (\hat{\beta}_{bj} - \hat{\beta}_j).$$

Stability, denoted by $sd(\hat{\beta}_j)$, was measured by the standard deviation of parameter estimates across replications and was computed as

$$sd(\hat{\beta}_j) = \sqrt{\frac{1}{999} \sum_{b=1}^{1000} \hat{\beta}^2_{bj}}.$$
Let $se(\hat{\beta}_{bj})$ denote the estimated standard error of parameter estimate $\hat{\beta}_{bj}$. Bias of the reported standard errors (BSE) was computed as

$$BSE = \frac{1}{1000} \sum_{h=1}^{1000} \left[ se(\hat{\beta}_{bj}) - se(\hat{\beta}_j) \right],$$

The bias and stability of parameter estimates and bias of the standard errors were only considered for parameters $\beta_2$ (a main effect that is influenced by the interaction effect $\beta_{23}$), $\beta_4$ (a main effect that is not influenced by the interaction effect $\beta_{23}$), and the interaction effect $\beta_{23}$.

### 3.2 Results

#### 3.2.1 Bias

Table 1 shows the most important results for bias for 40% missingness. For lower percentages of missingness, the pattern of the bias was similar but the absolute values were smaller. This confirms that for larger percentages of missingness, the imputation model becomes more important. The most important result is that incomplete-data methods MLID, MILC (AIC), and MICE (LOG) produced

---

**Table 1. Bias in the estimates of three logistic regression coefficients for six incomplete-data methods, three different levels of strength of three-variable associations ($\beta_{23} = 0, \beta_{23} = -0.8, \beta_{23} = -2$), two sample sizes (200, 1000), and 40% missingness.**

<table>
<thead>
<tr>
<th>N</th>
<th>$\beta_{23}$</th>
<th>RC</th>
<th>MLID</th>
<th>MILL</th>
<th>MILC (large)</th>
<th>MILC (AIC)</th>
<th>MICE (PMM)</th>
<th>MICE (LOG)</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>0</td>
<td>$\beta_2 = 1$</td>
<td>.040</td>
<td>.076</td>
<td>.082</td>
<td>.025</td>
<td>.051</td>
<td>.073</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_4 = 1$</td>
<td>.065</td>
<td>.067</td>
<td>.061</td>
<td>.041</td>
<td>.068</td>
<td>.061</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_{23} = 0$</td>
<td>.046</td>
<td>-.014</td>
<td>-.008</td>
<td>.031</td>
<td>.014</td>
<td>.009</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_{23} = -0.8$</td>
<td>.050</td>
<td>.094</td>
<td>.086</td>
<td>.040</td>
<td>.058</td>
<td>.058</td>
</tr>
<tr>
<td></td>
<td>-.8</td>
<td>$\beta_2 = 1.4$</td>
<td>.071</td>
<td>.057</td>
<td>.058</td>
<td>.042</td>
<td>.062</td>
<td>.057</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_4 = 1$</td>
<td>-.036</td>
<td>.010</td>
<td>.022</td>
<td>.258</td>
<td>.036</td>
<td>.310</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_{23} = -2$</td>
<td>.091</td>
<td>.103</td>
<td>.057</td>
<td>-.241</td>
<td>.057</td>
<td>-.357</td>
</tr>
<tr>
<td></td>
<td>-2</td>
<td>$\beta_2 = 2$</td>
<td>.058</td>
<td>.040</td>
<td>.036</td>
<td>.002</td>
<td>.055</td>
<td>.031</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_4 = 1$</td>
<td>-.144</td>
<td>-.074</td>
<td>-.006</td>
<td>.541</td>
<td>-.025</td>
<td>.733</td>
</tr>
<tr>
<td>1000</td>
<td>0</td>
<td>$\beta_2 = 1$</td>
<td>-.005</td>
<td>.010</td>
<td>.009</td>
<td>-.003</td>
<td>.011</td>
<td>.008</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_4 = 1$</td>
<td>.015</td>
<td>.013</td>
<td>.015</td>
<td>.012</td>
<td>.016</td>
<td>.014</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_{23} = 0$</td>
<td>-.019</td>
<td>-.004</td>
<td>-.006</td>
<td>-.001</td>
<td>-.006</td>
<td>-.004</td>
</tr>
<tr>
<td></td>
<td>-.8</td>
<td>$\beta_2 = 1.4$</td>
<td>.026</td>
<td>.036</td>
<td>.036</td>
<td>-.088</td>
<td>.036</td>
<td>-.136</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_4 = 1$</td>
<td>-.010</td>
<td>-.016</td>
<td>-.015</td>
<td>.205</td>
<td>.014</td>
<td>.302</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_{23} = -2$</td>
<td>.006</td>
<td>.033</td>
<td>.028</td>
<td>-.216</td>
<td>.031</td>
<td>.410</td>
</tr>
<tr>
<td></td>
<td>-2</td>
<td>$\beta_2 = 2$</td>
<td>.014</td>
<td>.011</td>
<td>.011</td>
<td>-.015</td>
<td>.012</td>
<td>-.013</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_4 = 1$</td>
<td>-.036</td>
<td>-.046</td>
<td>-.041</td>
<td>.457</td>
<td>-.056</td>
<td>.765</td>
</tr>
</tbody>
</table>

Note: N: sample size; $\beta_{23}$: strength of three-variable association; RC: regression coefficient. For MILC (AIC) the average number of classes indicated by AIC ranged from 2.8 ($N = 200$, $\beta_{23} = -2$) to 3.8 ($N = 1000$, $\beta_{23} = -2$), for MILC (Large) a constant number of 12 classes was used.

*Remarkable bias is printed in boldface.*
large bias in the estimates of $\beta_2$ and $\beta_{23}$ when there was an interaction effect in the data ($\beta_{23} \neq 0$), whereas estimates of $\beta_4$, a parameter not influenced by an interaction effect, showed much less bias. These results suggest that MILC (AIC) and MICE (LOG) have imputation models that are too parsimonious to pick up the three-way association in the data. Furthermore, the results suggest that MLID cannot handle very well the combination of a small sample size and a complex association. Seemingly, the asymptotic property of unbiased parameter estimates is not fully established under these circumstances. A second result is that MILL, which we used as a gold standard, produced similar bias or sometimes more bias (e.g., for $\beta_2$ in condition $\beta_{23} = -2$, $N = 200$) than MILC (Large) and MICE (PMM). A third result is that the bias was slightly larger for $N = 200$ than for $N = 1000$, which indicates that increased sampling variability may somewhat increase bias.

### 3.2.2 Stability

Table 2 shows the most important results for stability for 40% missingness. The most important result is that stability does not change dramatically across incomplete-data methods. MILC (AIC) was slightly more stable than MILC (Large), and MICE (LOG) is slightly more stable than MICE (PMM). This was expected because MILC (AIC) and MICE (LOG) are more parsimonious than MILC (Large) and MICE (PMM), respectively. The expected result that MILL would be less stable than MLID was not demonstrated. As expected, sample size had a positive effect on stability. For small samples ($N = 200$), the stability could be considered low, resulting in low power. For example,

<table>
<thead>
<tr>
<th>N</th>
<th>$\beta_{23}$</th>
<th>RC</th>
<th>MLID</th>
<th>MILL</th>
<th>MILC (large)</th>
<th>MILC (AIC)</th>
<th>MICE (PMM)</th>
<th>MICE (LOG)</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>0</td>
<td>$\beta_2 = 1$</td>
<td>.862</td>
<td>.845</td>
<td>.858</td>
<td>.600</td>
<td>.951</td>
<td>.738</td>
</tr>
<tr>
<td></td>
<td>$\beta_4 = 1$</td>
<td>.500</td>
<td>.496</td>
<td>.502</td>
<td>.445</td>
<td>.518</td>
<td>.501</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\beta_{23} = 0$</td>
<td>1.19</td>
<td>1.17</td>
<td>1.17</td>
<td>.840</td>
<td>1.23</td>
<td>.729</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\beta_{23} = -0.8$</td>
<td>.938</td>
<td>.931</td>
<td>.924</td>
<td>.679</td>
<td>1.01</td>
<td>.780</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\beta_2 = 1.4$</td>
<td>.505</td>
<td>.506</td>
<td>.510</td>
<td>.448</td>
<td>.533</td>
<td>.508</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\beta_4 = 1$</td>
<td>1.21</td>
<td>1.22</td>
<td>1.20</td>
<td>.862</td>
<td>1.26</td>
<td>.737</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\beta_{23} = -0.8$</td>
<td>.916</td>
<td>.956</td>
<td>.948</td>
<td>.748</td>
<td>1.02</td>
<td>.806</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>0</td>
<td>$\beta_2 = 1$</td>
<td>.444</td>
<td>.373</td>
<td>.370</td>
<td>.264</td>
<td>.380</td>
<td>.301</td>
</tr>
<tr>
<td></td>
<td>$\beta_4 = 1$</td>
<td>.203</td>
<td>.206</td>
<td>.206</td>
<td>.188</td>
<td>.206</td>
<td>.205</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\beta_{23} = 0$</td>
<td>.479</td>
<td>.515</td>
<td>.509</td>
<td>.362</td>
<td>.522</td>
<td>.306</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\beta_{23} = 1.4$</td>
<td>.365</td>
<td>.389</td>
<td>.384</td>
<td>.291</td>
<td>.392</td>
<td>.313</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\beta_4 = 1$</td>
<td>.205</td>
<td>.207</td>
<td>.208</td>
<td>.188</td>
<td>.207</td>
<td>.204</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\beta_{23} = -0.8$</td>
<td>.470</td>
<td>.507</td>
<td>.494</td>
<td>.376</td>
<td>.501</td>
<td>.310</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\beta_2 = 2$</td>
<td>.377</td>
<td>.407</td>
<td>.404</td>
<td>.342</td>
<td>.402</td>
<td>.306</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\beta_4 = 1$</td>
<td>.200</td>
<td>.205</td>
<td>.205</td>
<td>.185</td>
<td>.205</td>
<td>.197</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\beta_{23} = -2$</td>
<td>.482</td>
<td>.526</td>
<td>.517</td>
<td>.451</td>
<td>.514</td>
<td>.298</td>
<td></td>
</tr>
</tbody>
</table>

Note: N: sample size; $\beta_{23}$: strength of three-variable association; RC: regression coefficient.
the population value of $\beta_4$ was equal to 1, but in case $sd(\hat{\beta}_4) = .501$ (MILL, medium three-variable association), which is even one of the smaller standard deviations we found, one may expect to find estimates of $\beta_4$ between .02 and 1.98 (95% confidence interval). For large samples ($N = 1000$), the stability is much better. Percentage of missingness also had a negative effect on the stability. This can be expected because a larger percentage of missingness in fact means a reduction of the sample size.

### 3.2.3 Bias of the standard errors

Table 3 shows the most important results for bias of the standard errors for 20% missingness. Bias of the standard errors was smaller for $N = 1000$ than for $N = 200$. Bias of the standard errors was largest for the parameters associated with the three-variable association ($\beta_2$ and $\beta_{23}$) For $N = 200$, MILL and MILC (Large) had the smallest bias, whereas MILC (AIC) and MICE (LOG) overestimated the standard errors and MLID and MICE (PMM) underestimated the standard errors. For $N = 1000$, MILL, MILC (AIC), MILC (Large), and MICE (PMM) had the smallest bias, whereas MLID underestimated and MICE (LOG) overestimated the standard errors. This renders MILC (Large) as the most favorable incomplete-data method with respect to bias in standard errors.

**Table 3.** Bias in the standard errors of the estimates of three logistic regression coefficients for six incomplete-data methods, three different levels of strength of three-variable associations ($\beta_{23} = 0$, $\beta_{23} = -.8$, $\beta_{23} = -2$), two sample sizes (200, 1000), and 40% missingness.\(^a\)

<table>
<thead>
<tr>
<th>N</th>
<th>$\beta_{23}$</th>
<th>RC</th>
<th>MLID</th>
<th>MILL</th>
<th>MILC (large)</th>
<th>MILC (AIC)</th>
<th>MICE (PMM)</th>
<th>MICE (LOG)</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>0</td>
<td>$\beta_2 = 1$</td>
<td>-.040</td>
<td>-.044</td>
<td>-.044</td>
<td>.151</td>
<td>-.202</td>
<td>.072</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_4 = 1$</td>
<td>-.025</td>
<td>-.021</td>
<td>-.022</td>
<td>.015</td>
<td>-.052</td>
<td>-.021</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_{23} = 0$</td>
<td>-.092</td>
<td>.108</td>
<td>-.077</td>
<td>.174</td>
<td>-.219</td>
<td>.296</td>
</tr>
<tr>
<td></td>
<td>-.8</td>
<td>$\beta_2 = 1.4$</td>
<td>-.086</td>
<td>-.077</td>
<td>-.064</td>
<td>.116</td>
<td>-.214</td>
<td>.073</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_4 = 1$</td>
<td>-.027</td>
<td>-.017</td>
<td>-.018</td>
<td>.022</td>
<td>-.054</td>
<td>-.018</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_{23} = -.8$</td>
<td>-.092</td>
<td>-.077</td>
<td>-.062</td>
<td>.194</td>
<td>-.209</td>
<td>.315</td>
</tr>
<tr>
<td></td>
<td>-2</td>
<td>$\beta_2 = 2$</td>
<td>-.032</td>
<td>-.065</td>
<td>-.045</td>
<td>.080</td>
<td>-.199</td>
<td>.065</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_4 = 1$</td>
<td>-.010</td>
<td>-.015</td>
<td>-.017</td>
<td>.024</td>
<td>-.052</td>
<td>-.012</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_{23} = -2$</td>
<td>-.074</td>
<td>-.064</td>
<td>-.054</td>
<td>.164</td>
<td>-.193</td>
<td>.292</td>
</tr>
<tr>
<td>1000</td>
<td>0</td>
<td>$\beta_2 = 1$</td>
<td>-.001</td>
<td>-.015</td>
<td>-.011</td>
<td>.065</td>
<td>-.060</td>
<td>.038</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_4 = 1$</td>
<td>-.006</td>
<td>-.005</td>
<td>-.005</td>
<td>.008</td>
<td>-.010</td>
<td>-.005</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_{23} = 0$</td>
<td>-.015</td>
<td>-.030</td>
<td>-.025</td>
<td>.088</td>
<td>-.086</td>
<td>.130</td>
</tr>
<tr>
<td></td>
<td>-.8</td>
<td>$\beta_2 = 1.4$</td>
<td>-.017</td>
<td>-.023</td>
<td>-.021</td>
<td>.041</td>
<td>-.067</td>
<td>.029</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_4 = 1$</td>
<td>-.007</td>
<td>-.006</td>
<td>-.007</td>
<td>.009</td>
<td>-.011</td>
<td>-.004</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_{23} = -.8$</td>
<td>-.005</td>
<td>-.015</td>
<td>-.005</td>
<td>.070</td>
<td>-.059</td>
<td>.126</td>
</tr>
<tr>
<td></td>
<td>-2</td>
<td>$\beta_2 = 2$</td>
<td>-.012</td>
<td>.040</td>
<td>-.036</td>
<td>.080</td>
<td>-.074</td>
<td>.032</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_4 = 1$</td>
<td>.002</td>
<td>-.002</td>
<td>-.000</td>
<td>.024</td>
<td>-.006</td>
<td>.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_{23} = -2$</td>
<td>.002</td>
<td>.028</td>
<td>-.028</td>
<td>.164</td>
<td>-.068</td>
<td>.133</td>
</tr>
</tbody>
</table>

Note: N = sample size; $\beta_{23}$ = strength of three-variable association; RC = regression coefficient.

\(^a\)Remarkable bias is printed in boldface.
4 Study 2: bias, stability, and bias in standard errors produced by MILC, MICE and complete-case analysis for a larger number of trichotomous variables

In Study 2, we compared incomplete-data methods MILC (AIC), MILC (Large, $K=33$), MICE (PMM), and MICE (LOG) to complete-case analysis, on bias of the parameter estimates, stability of the parameter estimates, and bias of the standard errors. Benchmarks MLID and MILL could no longer be used because the number of variables (11) was too large. The main question for Study 2 was whether MILC and MICE would also work for polytomous categorical data and for large numbers of possible response patterns. In Study 1, the number of possible response patterns was $2^6=64$, whereas in Study 2, the number of possible response patterns was increased to $3^{11}=177,147$. The main objective for the design of Study 2 is that the associations among the variables need to be complex, to test whether the incomplete-data methods can pick up the associations correctly.

4.1 Method

4.1.1 General set-up

In Study 2, the population model from which the complete data sets were sampled contained eight trichotomous predictor variables ($Y_1, \ldots, Y_8$) and three trichotomous outcome variables ($Y_9, Y_{10}$, and $Y_{11}$). The categories were coded $-1, 0, 1$. The associations among the 11 are described by a path model for categorical data\textsuperscript{46} containing one-, two-, and three-way associations (see Figure 1 for a graphical representation and Appendix A for the chosen parameter values).

Variables $Y_1, Y_3, Y_4,$ and $Y_{11}$ had missing values; the other variables were completely observed. The missingness mechanism was MAR, and rather complex. For $Y_1$ and $Y_3$, the missingness depended on $Y_2$ and $Y_9$. Let $R$ indicate whether (score 1) or not (score 0) a score is observed. Both for $Y_1$ and $Y_3$, the logit of $R$ was $\text{logit}(R) = -5.06 - 2 \cdot Y_9 + 3 \cdot Y_2$, resulting in approximately 20% missing values for each variable. Similarly, for $Y_4$ and $Y_{11}$, the missingness depended on $Y_7$ and $Y_9$. Both for $Y_4$ and $Y_{11}$, $\text{logit}(R) = -5.50 + 3 \cdot Y_9 - 1.5 \cdot Y_7$, also resulting in approximately 20% missing values for each variable. This procedure kept the total percentage of missingness constant at $4/11 \cdot 20\% + 7/11 \cdot 0\% = 7.27\%$.

For each incomplete data set, the multiple imputation methods created $m=5$ completed data sets. For complete-case analysis, a complete data set was obtained by simply deleting every row that contained at least one missing value.

The substantive model was an adjacent category ordinal logit model\textsuperscript{35} for outcome variable $Y_9$ containing $Y_1, Y_3, Y_4,$ and $Y_{11}$ as predictors. The logit equation has the form

$$\text{logit}(Y_9 = j \mid Y_9 = j-1 \text{ or } Y_9 = j) = \beta_{0j} + \beta_1 Y_1 + \beta_2 Y_2 + \beta_3 Y_3 + \beta_4 Y_4 + \beta_{12} Y_1 Y_2,$$

for $j=1, 2$. Note that the substantive model is part of the population model (Figure 1) and includes the main effects of the predictors of $Y_9$, and the interaction effect of $Y_1$ and $Y_2$. The latter implies a three-variable association among $Y_1, Y_2,$ and $Y_9$.

Three software packages were used for data generation, incomplete-data handling, and estimating the substantive model. Complete and incomplete data were generated by LEM.\textsuperscript{20} The imputation phase of MILC (Large) and MICE (PMM) was performed using Latent GOLD and the R package MICE, respectively. Latent GOLD was used to estimate the substantive model for by MILC (Large) and MICE (PMM), using the completed datasets, and for MLID and complete-case analysis.
4.1.2 Design

We only varied sample size and incomplete-data method. Sample size had two levels: medium ($N = 500$) and large ($N = 1000$); incomplete-data method had five levels: MILC (AIC), MILC (Large), MICE (LOG), MICE (PMM), and complete-case analysis. This yields a $5 \times 2$ design. The outcome variables were equivalent to those in Study 1 (bias, stability, and bias of standard errors).

4.2 Results

4.2.1 Bias

Table 4 shows the bias for $\beta_2$, the main effect of a predictor that is also involved in the interaction effect; $\beta_3$, the main effect of a predictor not involved in the interaction effect; and $\beta_{12}$, the interaction effect itself. The most important result is that MICE (PMM) and MICE (LOG) produced relatively large bias in the estimates of $\beta_3$ and $\beta_{12}$, suggesting that the imputation models of MICE (LOG) and MICE (PMM) do not correctly pick up the three-way association in the data. Furthermore, complete-case analysis produced very large bias in the estimates of $\beta_2$ and $\beta_{12}$, confirming that
Table 4. Bias in the estimates of three logistic regression coefficients for five incomplete-data methods, two sample sizes (500, 1000), and 20% missingness on four variables.\textsuperscript{a}

<table>
<thead>
<tr>
<th>N</th>
<th>RC</th>
<th>CC</th>
<th>MILC (large)</th>
<th>MILC (AIC)</th>
<th>MICE (PMM)</th>
<th>MICE (LOG)</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>$\beta_2 = -.45$</td>
<td>.504</td>
<td>-.033</td>
<td>-.028</td>
<td>-.035</td>
<td>-.036</td>
</tr>
<tr>
<td></td>
<td>$\beta_3 = .5$</td>
<td>.017</td>
<td>.001</td>
<td>-.002</td>
<td>.054</td>
<td>.051</td>
</tr>
<tr>
<td></td>
<td>$\beta_{12} = .45$</td>
<td>-.110</td>
<td>-.068</td>
<td>-.066</td>
<td>-.114</td>
<td>-.113</td>
</tr>
<tr>
<td>1000</td>
<td>$\beta_2 = -.45$</td>
<td>.501</td>
<td>-.027</td>
<td>-.025</td>
<td>-.033</td>
<td>-.035</td>
</tr>
<tr>
<td></td>
<td>$\beta_3 = .5$</td>
<td>.019</td>
<td>.001</td>
<td>-.003</td>
<td>.053</td>
<td>.046</td>
</tr>
<tr>
<td></td>
<td>$\beta_{12} = .45$</td>
<td>-.113</td>
<td>-.061</td>
<td>-.061</td>
<td>-.116</td>
<td>-.114</td>
</tr>
</tbody>
</table>

Note: N: sample size; RC: regression coefficient; CC: complete-case analysis.
\textsuperscript{a} Remarkable bias is printed in boldface.

Table 5. Stability of the estimates of three logistic regression coefficients for five incomplete-data methods, two sample sizes (500, 1000), and 20% missingness on four variables

<table>
<thead>
<tr>
<th>N</th>
<th>RC</th>
<th>CC</th>
<th>MILC (large)</th>
<th>MILC (AIC)</th>
<th>MICE (PMM)</th>
<th>MICE (LOG)</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>$\beta_2 = -.45$</td>
<td>.101</td>
<td>.290</td>
<td>.288</td>
<td>.293</td>
<td>.292</td>
</tr>
<tr>
<td></td>
<td>$\beta_3 = .5$</td>
<td>.215</td>
<td>.218</td>
<td>.215</td>
<td>.240</td>
<td>.240</td>
</tr>
<tr>
<td></td>
<td>$\beta_{12} = .45$</td>
<td>.146</td>
<td>.159</td>
<td>.157</td>
<td>.127</td>
<td>.130</td>
</tr>
<tr>
<td>1000</td>
<td>$\beta_2 = -.45$</td>
<td>.069</td>
<td>.272</td>
<td>.271</td>
<td>.275</td>
<td>.275</td>
</tr>
<tr>
<td></td>
<td>$\beta_3 = .5$</td>
<td>.224</td>
<td>.222</td>
<td>.220</td>
<td>.247</td>
<td>.246</td>
</tr>
<tr>
<td></td>
<td>$\beta_{12} = .45$</td>
<td>.134</td>
<td>.162</td>
<td>.161</td>
<td>.134</td>
<td>.137</td>
</tr>
</tbody>
</table>

Note: N: sample size; RC: regression coefficient; CC: complete-case analysis.

complete-case analysis leads to biased results when the data are MAR. MILC (AIC) and MILC (Large) had a similar performance in terms of bias.

4.2.2 Stability
Table 5 shows the stability of $\beta_2$, $\beta_3$, and $\beta_{12}$. The most important result is that (almost) unbiased parameter estimates (see Table 4) showed similar stability across methods, whereas biased parameter estimates tended to be either more stable or more unstable. This effect was clearer for $N = 500$ than for $N = 1000$. For example, for the estimate of $\beta_2$, MILC (Large), MILC (AIC), MICE (PMM), and MICE (LOG) show similar bias and similar stability of parameter estimates. However, complete-case analysis overestimated $\beta_2$ and this estimate was too stable, whereas MICE (PMM) and MICE (LOG) overestimated $\beta_3$ and this estimate was too unstable.

4.2.3 Bias of the standard errors
Table 6 shows the bias of the standard errors of $\beta_2$, $\beta_3$, and $\beta_{12}$. Bias of the standard errors was smaller for $N = 1000$ than for $N = 500$. The multiple imputation methods yielded similar upward bias in their standard errors, whereas complete-cases analysis tended to yield a larger overestimation of the standard errors. Bias was largest for the standard error of the interaction effect ($\beta_{12}$).
We applied the most promising variants of MILC and MICE (i.e., MILC (Large, K = 12) and MICE (PMM)), complete-case analysis, and MLID to data from the Investigators of Projective Services Project for Older Persons\textsuperscript{47}, which have been discussed and analyzed earlier by Fuchs\textsuperscript{48} to illustrate the MLID approach. The data set contains the scores of 164 patients on six dichotomous variables (Table 7). One patient had a missing value on the physical status, 33 had a missing value for mental status, and 29 respondents had a missing value on both physical and mental status. The question of interest is whether the unexpected negative association between treatment and survival disappears when controlling for age, gender, physical status, and mental status.

The substantive model predicts survival by the main effects of all variables, plus the interaction effect of mental status, $Y_1$, and physical status, $Y_2$. We defined the following regression model,

$$\text{logit}(Y_6) = \beta_0 + \beta_1 Y_1 + \beta_2 Y_2 + \beta_3 Y_3 + \beta_4 Y_4 + \beta_5 Y_5 + \beta_{12} Y_1 Y_2.$$  

Contrary to Fuchs, we chose to include the interaction between mental status and physical status because we were interested in whether the imputation methods yielded similar results to MLID in a model containing a higher-order association. Once the data had been imputed using MILC (Large)

\begin{center}
\begin{table}
\caption{Bias in the standard errors of the estimates of three logistic regression coefficients for five incomplete-data methods, two sample sizes (500, 1000), and 20% missingness on four variables\textsuperscript{a}}
\label{table6}
\centering
\begin{tabular}{llcccccc}
\hline
\text{Incomplete-data method} & & & \text{MILC (large)} & & \text{MILC (AIC)} & & \text{MICE (PMM)} & & \text{MICE (LOG)} \\
\hline
\text{N} & \text{RC} & \text{CC} & \text{MILC (large)} & \text{MILC (AIC)} & \text{MICE (PMM)} & \text{MICE (LOG)} \\
\hline
500 & $\beta_2 = -.45$ & .126 & .080 & .080 & .082 & .081 \\
& $\beta_3 = .5$ & .152 & .081 & .083 & .088 & .084 \\
& $\beta_{12} = .45$ & .156 & .100 & .102 & .106 & .102 \\
1000 & $\beta_2 = -.45$ & .089 & .058 & .058 & .058 & .057 \\
& $\beta_3 = .5$ & .086 & .061 & .062 & .062 & .059 \\
& $\beta_{12} = .45$ & .111 & .076 & .077 & .075 & .072 \\
\hline
\end{tabular}
\end{table}
\end{center}

Note: N: sample size; RC: regression coefficient; CC: complete-case analysis.
\textsuperscript{a}Remarkable bias is printed in boldface.

\begin{center}
\begin{table}
\caption{Information on the variables of the Projective Services Project for older persons\textsuperscript{47}}
\label{table7}
\centering
\begin{tabular}{lll}
\hline
\text{Variable} & \text{Levels} & \text{Code} \\
\hline
Mental status & Poor, good & $Y_1$ \\
Physical status & Poor, good & $Y_2$ \\
Age & less than 75, over 75 & $Y_3$ \\
Group membership & experiment, control & $Y_4$ \\
Sex & male, female & $Y_5$ \\
Survival status & Deceased, survived & $Y_6$ \\
\hline
\end{tabular}
\end{table}
\end{center}
and MICE (PMM), the substantive model defined in equation (7) was estimated. We also estimated the logistic regression model using MLID (as a benchmark) and complete-case analysis. Schafer’s notion on the number of variables is of no concern in this analysis because the substantive model and the imputation model are identical; both models include all available variables. Therefore, the performance of MILC (Large), MICE (PMM), and complete-case analysis was assessed by comparing them to MLID (Table 8).

For all incomplete-data methods, only age (β3, negative effect) had a significant effect on survival status. The fact that all other effects were not statistically significant may be due to the small sample size. Nevertheless, it remains interesting to compare the parameter estimates across incomplete-data methods. Table 8 shows that the estimates yielded by MILC (Large) were very similar to MLID, for all parameters. MICE (PMM) produced estimates of β3, β4 and β5 that were very similar to MLID, but yielded relatively large differences for parameters β1, β2, and β12. Complete-case analysis produced rather large differences for the estimates of β2, β4, β5 and β12. MLID, MILC (Large), and MICE (PMM) did not have large differences in the estimated standard errors. However, complete-case analysis yielded relatively large standard errors for parameters β3, β4, and β5, compared to MLID.

6 Discussion

The aim of this paper was to investigate which incomplete-data method for categorical data should be recommended to practitioners. We assessed the performance of MILC and MICE with regard to three criteria, relative to MLID, MILL, and complete-case analysis. Based on the theoretical discussion, Study 1, and Study 2, MILC (Large) appears to be the incomplete-data method that meets the three criteria to the greatest extent. The other incomplete-data methods have one or more features that make them suboptimal. MILL cannot handle more than a few variables, MLID does not allow for the use of small substantive model as it can affect the MAR assumption, and may yield biased parameter estimates for a complex association in case of a small sample size. While in Study 1 MICE (PMM) performed rather well, Study 2 showed that MICE (PMM) may yield biased parameter estimates when the number of possible data pattern is large, especially when the sample size is small. MILC (AIC) and MICE (LOG) may fail to capture higher-order associations in the data, which yields parameter estimates with an unacceptably high bias.

Table 8. Estimated logistic regression coefficients using MLID, MILC (Large), and MICE (PMM)\(^a\)

<table>
<thead>
<tr>
<th>RC</th>
<th>Est.</th>
<th>SE</th>
<th>Est.</th>
<th>SE</th>
<th>Est.</th>
<th>SE</th>
<th>Est.</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>β1</td>
<td>3.175</td>
<td>2.188</td>
<td>2.844</td>
<td>2.070</td>
<td>3.777</td>
<td>1.984</td>
<td><strong>2.271</strong></td>
<td>1.815</td>
</tr>
<tr>
<td>β2</td>
<td>2.614</td>
<td>2.240</td>
<td><strong>1.816</strong></td>
<td>2.180</td>
<td>3.162</td>
<td>2.102</td>
<td><strong>1.439</strong></td>
<td>1.979</td>
</tr>
<tr>
<td>β3</td>
<td>-1.417</td>
<td>.431</td>
<td>-1.568</td>
<td>.526</td>
<td>-1.380</td>
<td>.422</td>
<td>-1.426</td>
<td>.434</td>
</tr>
<tr>
<td>β4</td>
<td>.459</td>
<td>.394</td>
<td>.176</td>
<td>.481</td>
<td>.496</td>
<td>.392</td>
<td>.475</td>
<td>.384</td>
</tr>
<tr>
<td>β5</td>
<td>-.506</td>
<td>.417</td>
<td>-.281</td>
<td>.525</td>
<td>-.420</td>
<td>.437</td>
<td>-.583</td>
<td>.410</td>
</tr>
<tr>
<td>β12</td>
<td>-1.017</td>
<td>1.269</td>
<td><strong>-1.629</strong></td>
<td>1.217</td>
<td>-1.283</td>
<td>1.168</td>
<td><strong>-1.382</strong></td>
<td>1.090</td>
</tr>
</tbody>
</table>

Note: RC: regression coefficient; Est: estimate; SE: standard error.

\(^a\)Remarkable bias is printed in boldface

van der Palm et al. 771
In Study 2 it was demonstrated that complete-case analysis yields very large bias in the parameter estimates and a loss of power due to inflated standard errors. The findings in the real-data example were consistent with these results.

A remaining issue with MILC is that there is not yet a guideline indicating the minimum number of required latent classes. The simulation study showed that over fit does not seem to be a problem, which was also argued by Vermunt et al., so one can always resort to estimating a latent class model with many classes. However, having a minimum of required latent classes would greatly facilitate the use of MILC because estimating latent class models with 40, 50, or 60 latent classes can be very time consuming. We showed that in case of a small table AIC is not a good criterion because the number of classes is too low; for a large table MILC (AIC) yielded good results. A heuristic rule may be to use as many classes as there are categories in the data. For example, for a data set consisting of 10 variables with three response categories and 5 dichotomous variables, the number of latent classes would be $10 \times 3 + 5 \times 2 = 40$ latent classes. Whether this heuristic rule is reasonable should be investigated in future research.

An additional comment should be made for MICE (LOG), as it may have been presented too negatively. The problem of MICE (LOG) is that in the default setting, interaction effects are not included in the conditional models. As a result, the imputation model may be too parsimonious yielding biased parameter estimates. Further research would be required to investigate whether this method is able to produce unbiased results if the conditional model included higher-order interactions.

Lastly, we note that each incomplete-data method had to be applied using a different software package, as there is no package available that applies all of the methods. Further research is warranted to investigate the potential differences between implementations of MILC and MICE across software packages.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Daniël W van der Palm is supported by a grant from the Netherlands Organisation for Scientific Research (NWO 400-08-234).

References

6. Schafer JL. *Analysis of incomplete multivariate data*.
36. *StataCorp*. *Stata statistical software: Release 12*. College Station, TX: StataCorp. 2011..

**Appendix A**

Tables A1, A2, and A3 show the parameter values describing the population model in Study 2 (Figure 1). All variables had three nominal response categories. Because dummy coding was used, the effect of the first category was zero (not displayed). For all two-way and three-way interactions only a single value is shown because the associations were defined to be ordinal (linear-by-linear).
Table A1. Loglinear parameters describing \( P(Y_1, Y_2, Y_3, Y_4, Y_5) \)

<table>
<thead>
<tr>
<th>( Y_1 )</th>
<th>( Y_1 Y_2 )</th>
<th>( Y_2 Y_4 )</th>
<th>( Y_1 Y_4 Y_5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(.15,.0)</td>
<td>.3</td>
<td>.55</td>
<td></td>
</tr>
<tr>
<td>(.25,.05)</td>
<td>-.4</td>
<td>-.36</td>
<td>-.25</td>
</tr>
<tr>
<td>(-.15, -.15)</td>
<td>-.2</td>
<td>-.15</td>
<td>.55</td>
</tr>
<tr>
<td>(.05,.25)</td>
<td>.5</td>
<td>-.05</td>
<td></td>
</tr>
<tr>
<td>(-.25, -.05)</td>
<td>-.3</td>
<td>.3</td>
<td></td>
</tr>
</tbody>
</table>

Table A2. Loglinear parameters describing \( P(Y_6, Y_7, Y_8) \)

<table>
<thead>
<tr>
<th>( Y_6 )</th>
<th>( Y_6 Y_7 )</th>
<th>( Y_6 Y_8 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(.15,.75)</td>
<td>.32</td>
<td>-.4</td>
</tr>
<tr>
<td>(.25,1.25)</td>
<td>.25</td>
<td>.24</td>
</tr>
<tr>
<td>(.1,.5)</td>
<td>.24</td>
<td>-.4</td>
</tr>
</tbody>
</table>

Table A3. Logistic regression parameters

| \( Y_9 \) \( | Y_1 = -.3 \) | \( Y_9 \) \( | Y_4 = .22 \) | \( Y_9 \) \( | Y_6 = -.38 \) | \( Y_9 \) \( | Y_9 = -.14 \) |
|----------|----------------|----------------|----------------|----------------|
| \( Y_9 \) \( | Y_2 = -.45 \) | \( Y_9 \) \( | Y_5 = .32 \) | \( Y_9 \) \( | Y_6 = .34 \) | \( Y_9 \) \( | Y_{11} = .35 \) |
| \( Y_9 \) \( | Y_3 = .5 \)  | \( Y_9 \) \( | Y_6 = .42 \) | \( Y_9 \) \( | Y_{10} = .22 \) | \( Y_9 \) \( | Y_{11} = -.15 \) |
| \( Y_9 \) \( | Y_1 Y_2 = .45 \) | \( Y_9 \) \( | Y_7 = -.38 \) | \( Y_9 \) \( | Y_6 Y_7 = .4 \) | \( Y_9 \) \( | Y_9 = -.14 \) |

Table A1 shows the log linear parameters describing \( P(Y_1, Y_2, Y_3, Y_4, Y_5) \), Table A2 shows the log linear parameters describing \( P(Y_6, Y_7, Y_8) \), and Table A3 shows the logistic regression parameters relating predictor variables \( Y_1, Y_2, Y_3, Y_4, Y_5, Y_6, Y_7, Y_8 \) to outcome variables \( Y_9, Y_{10}, Y_{11} \).