Journal of Classification 34:327-359 (2017)
DOI: 10.1007/s00357-017-9220-3

Handling Missing Data in Item Response Theory. Assessing the Accuracy of a Multiple Imputation Procedure Based on Latent Class Analysis

Isabella Sulis

Università degli Studi di Cagliari, Italy

Mariano Porcu

Università degli Studi di Cagliari, Italy

Abstract: A critical issue in analyzing multi-item scales is missing data treatment. Previous studies on this topic in the framework of item response theory have shown that imputation procedures are in general associated with more accurate estimates of item location and discrimination parameters under several missing data generating mechanisms. This paper proposes a model-based multiple imputation procedure for multiple categorical items (dichotomous, multinomial or Likert-type) which relies on the results of latent class analysis to impute missing item responses. The effectiveness of the proposed technique is assessed in the estimation of item response theory parameters using a range of *ad hoc* measures. The accuracy of the method is assessed with respect to other single and multiple imputation procedures, under different missing data generating mechanisms and different rate of missingness (5% to 30%). The simulation results indicate that the proposed technique performs satisfactorily under all conditions and has the greatest potential with severe rates of missingness and under non ignorable missing data mechanisms. The method was implemented in R code with a function that calls scripts from a latent class analysis routine.

Keywords: Item response theory; Multiple imputation analysis; Latent class analysis; Missingness; Accuracy measures.

The authors would like to thank the editor and the anonymous reviewers for their helpful and useful comments that greatly contributed in improving the final version of the paper.

Corresponding Author's Address: I. Sulis, Dipartimento di Scienze Sociali e delle Istituzioni, Università degli Studi di Cagliari, Tel. +39 0706753543, Fax. +39 0706753760, email: isulis@ unica.it.

Published online: 22 March 2017

1. Introduction

Item non-response is a frequent issue in item response theory (IRT) studies (Baker and Kim, 2004), where categorical items (also known as multi-item scales) are used to operationalize a latent trait of interest: e.g. in surveys designed to measure students' competencies in specific areas, people's opinions, attitudes, abilities or psychological constructs. The IRT framework links a person's responses to categorical items to an underlying continuous latent trait defining the probability that a certain category of an item will be selected as a function of the item itself and of a person's latent trait value. More specifically, IRT investigates how the probability of responses to an item varies as function of (i) the item-category position along the latent trait (item-category location parameter), (ii) the item capability to discriminate between individuals with different latent trait values (item discrimination parameter), and (iii) the individual's intensity of the latent trait (person parameter). Such probability is usually modelled using a logistic distribution. The main advantage of the IRT modeling approach is that item-category location parameters and person parameters are measured on the same metric. The probability to endorse a category of response in an item is positively related to the person parameter (θ) and the item discrimination parameter (λ) and negatively related to the item-category location (β) parameter. Many extensions of the approach have been advanced in the literature (Baker and Kim, 2004). The main differences across them is in the possibility to impose constraints to the item-category parameters, item slopes, and in the way the logistic function for multinomial responses (Agresti, 2002) is specified (Baker and Kim, 2004).

There are different types of missing responses that can be observed in the analysis of multi-item scales. It can happen that respondents skip one or more items unintentionally or that they do not have enough time to fill in all the responses. Or, it may be the case that respondents simply do not know how to answer, they do not have a clear awareness of what has been asked in the question, or that they do not want to report their opinion on a specific topic.

The typology of missing data generating processes, and the related implication thereof in terms of the reliability of the inferential results, seems to be strictly linked to the variety of reasons for missingness. See Schafer and Graham (2002), Sijtsma and Van Der Ark (2003), Enders (2004) and Finch (2008) for an exhaustive discussion on the topic.

In multi-item scales it is a common practice to handle missing values by filling in the empty holes in the data matrix with plausible values which are generated on the basis of deterministic or stochastic approaches (Rubin, 1976). The main rationale for imputation approaches in IRT framework is that the *built in* linking mechanism (Edelen and Reeve, 2007) at the basis of IRT models ensures that the set of items which define a scale of measurement for the underlying latent trait are calibrated to the same scale. This makes observed responses informative of non responses. Furthermore, imputation procedures are generally preferred to other missing data handling methods since they make it possible to proceed in further analysis with a complete data matrix.

The main strength of these imputation procedures is that the missing data problem is established before starting any analysis and standard statistical tools are used on data sets which contain imputed values instead of empty observations.

Multiple imputation procedures differ from single imputation methods because they generate more plausible values for each missing value, creating multiple versions of the same data set which can be analyzed separately (Rubin, 1976). This way of handling missing information takes into account the uncertainty related to the unknown real values while summarizing the results observed on multiple data sets in a single overall statement; this is the characteristic that makes this class of imputation methods more appealing in comparison to the others.

This paper discusses a multiple imputation procedure which relies on latent class analysis (LCA) for categorical items (dichotomous, multinomial or Likert-type) to handle with missing data in multi-item Likert scales.

The effectiveness of the proposed procedure was determined on two data sets: a multi-item Likert-type scale often used in surveys on students' evaluation of teaching and a multi-item Likert-type scale that it is used in the Progress in International Reading Literacy Study – PIRLS – survey 2011 (Mullis et al., 2012) for measuring students'attitude towards reading: in both simulation studies the observations were set as missing according to several missing data schemes. The study aimed to assess under which conditions the proposed procedure will have greater efficiency in the framework of IRT than do other missing data imputation methods, i.e. those that are chosen on the basis of their effectiveness in similar studies or/and due to their easy applicability for non-practitioners. Specifically, these other methods include, among the others, Multiple Imputation based on Multivariate Normal Distribution (MI), Multiple Imputation by Stochastic Regression (MISR), Multiple Imputation by Chain Equation (MICE) and Correct Mean Substitution (CMS) (Raaijmakers, 1999; Vermunt et al., 2008; Sulis and Porcu, 2008; Sulis, 2013).

Two main tasks have been simultaneously pursued in the study: (i) to validate the effectiveness of the proposed procedure in the estimation of IRT parameters under different missing data generating processes and with increasing rates of missingness; (ii) to evaluate the accuracy of the proposed methods compared to other widely used imputation procedures.

The structure of this paper is as follows. Section 2 presents a discussion of the missing data generating processes in the framework of IRT. Section 3 examines the rationale behind multiple imputation and the justification for adopting the proposed procedure in the framework of IRT. Section 4 is a detailed discussion on how the procedure works and briefly introduces the other four imputation procedures that will be adopted for comparative purposes. Section 5 describes the simulation study, advances a variety of accuracy measures to compare the effectiveness of the procedures, and presents the main results therein. The main findings which arise from the analysis are discussed in Section 6. The functions implemented in order to use the procedure with a data matrix of categorical (Likert-type) items and to simulate data affected by missingness under different missing data generating processes were implemented in the R language and are available in the supplementary online materials. The Tables containing detailed results of the simulation study are listed in the supplementary online materials.

2. Missing Data Classification. Focus on IRT Models

The method chosen to deal with missing information may cause bias, inefficiency or both in the estimation of key parameters, depending on whether or not the process which generates missing values can be ignored as well as on the treatment of the rate of missing observations (Rubin, 1976; Schafer, 1997). Rubin (1976) defines a taxonomy of missing values according to the process which generates unobserved responses in a data matrix $\boldsymbol{Y}$. Let's denote Y_o as the observed values of the data matrix and Y_m as the missing one. Define as $\boldsymbol{R}$ a missingness matrix composed of J ($j = 1, \ldots, J$) dummy variables, where each R_{ij} takes a value of 1 if the observation i ($i = 1, \ldots, n$) is missing and 0 otherwise. An analysis of the conditional distribution of $\boldsymbol{R}$ given $\boldsymbol{Y}$, allows us to identify the missing data generating process: *Missing Completely at Random* (MCAR), *Missing at Random* (MAR) or *Missing not at Random* (MNAR) (Rubin, 1976). Unobserved responses are MCAR if the probability of observing a missing value depends solely on the probability distribution of $\boldsymbol{R}$ (i.e., it is not dependent on the observed and/or the missing values): the $P(\boldsymbol{R}|Y_o, Y_m) = P(\boldsymbol{R})$. Under MCAR conditions, we can say that no particular causes are related to missingness. If we assume there to be a MCAR process, deleting any unit with incomplete values from the analysis, i.e. by performing a so called *Complete Case Analysis* (CCA), in general should not bias the final results even though the reduction of sample size causes a loss in efficiency (Schafer and Graham, 2002). Missing data are considered MAR when their probability distribution depends only on observed data $P(\boldsymbol{R}|Y_o, Y_m) = P(\boldsymbol{R}|Y_o)$. Thus, missing responses are usually predictable using Y_o (Little and Rubin, 2002; Schafer and Graham, 2002). Under MAR using a CCA can have different consequences depending on the parameter of interest: e.g. regression coefficients estimated using CCA are in general unbiased, whereas the parameters of the marginal distribution are in general biased, as well as the correlation coefficient between two variables (Schafer and Graham, 2002). Finally, the missing process is said to be MNAR if the probability of observing a missing value depends on the observed and unobserved units $P(\boldsymbol{R}|Y_o, Y_m) = P(\boldsymbol{R}|Y_o, Y_m)$; thus the missing data process is not ignorable and a CCA will produce a bias in the estimates of the parameters. Under MNAR unobserved values are not predictable using classic imputation methods on the basis of the observed units (Little and Rubin, 2002; Schafer and Graham, 2002). An appealing aspect for using imputation procedures in IRT is that items in the same scale share a certain degree of homogeneity because they are supposed to measure different segments of an underlying unidimensional latent trait.

In the IRT framework, the missing data process should be considered MCAR if the propensity to observe a missing value in an item is unrelated (i) to the value of the item itself and to the values of other items, (ii) to the latent trait values, and (iii) to any other measured variables in the analysis (Little and Rubin, 2002; Sijtsma and Van Der Ark, 2003; Enders, 2004). If missing observations in an item are related to other variables such as respondents' characteristics or responses to another item, the missing data process is said to be MAR. This is what occurs, for example, in a survey on students' evaluation of teaching (the first data on which our procedure has been validated) if the propensity for missing data depends on other student-related variables, as for instance to belong to groups of students with different levels of achievement or with different levels of interest towards the discipline (Sijtsma and Van Der Ark, 2003; Sulis and Porcu, 2008; Baraldi and Enders, 2010). Lastly, whenever the probability to observe a missing value is directly related to the latent trait values the mechanism is MNAR. In the survey of students' evaluation of teaching this is observed if students with different values of the satisfaction with respect to the university teaching have different propensity to skip responses to the items.

3. Rationale for Multiple Imputation

The debate on the effectiveness of *ad hoc* missing data imputation methods for multi-item scales has increased over the last few decades (Bernaards and Sijtsma, 1999; Raaijmakers, 1999; Huisman, 1999; Sijtsma and Van Der Ark, 2003; Enders, 2004; Finch, 2008, 2011; Carpita and Manisera, 2011). Many deterministic imputation methods have been advanced in IRT framework; they replace missing observations of a specific item with values set as a weighted or unweighted function of the responses of the person to the other items or/and as weighted or unweighted function of the responses provided to item affected by missingness by the other respondents. Other methods impute values on the basis of the responses provided for the item by individuals with similar response patterns (who, thereby act as donors). For an overview of possible options and their potentiality, we refer the interested reader to, among others, Raaijmakers (1999), Sijtsma and Van Der Ark (2003), Finch (2008) and Carpita and Manisera (2011). We can assert that in an IRT framework, discarding all partially observed units is not generally recommended (the default solution automatically adopted by many statistical packages) even when the missing data mechanism is ignorable (i.e., MCAR and MAR), whereas there is a general agreement on considering it a more efficient solution to impute the partially observed records with plausible values (Little and Rubin, 2002). As has been highlighted by many authors (Raaijmakers, 1999; Sijtsma and Van Der Ark, 2003; Bernaards and Sijtsma, 1999), deterministic imputation procedures (e.g. relative mean substitution), based on the weighting function of item and person responses can be considered valid alternatives to model-based approaches when the missing data mechanism is ignorable, when the rate of missingness is trivial, and where there is lack of expertise in implementing or in dealing with more complex procedures. This is the main reason why single imputation procedures based on weighted methods have been widely applied in IRT literature (Raaijmakers, 1999; Bernaards and Sijtsma, 1999; Sijtsma and Van Der Ark, 2003; Finch, 2008). Simulation studies highlight that they are usually superior to listwise deletion, mean imputation, random imputation and other *hot-deck* methods that fill in missing values with values from observed respondents. Differences among the statistical performances of the above mentioned imputation methods decrease as the percentage of missing values decreases, as the sample size increases, and as the level of association between variables decreases (Raaijmakers, 1999).

The value added in using a Multiple Imputation Analysis (MIA) (Little and Rubin, 2002) to deal with missingness in data analysis is that the method takes account of the uncertainty related to the unknown real values by imputing M plausible values for each unobserved response in the data set. In this way, the M imputed versions of the data set are identical for the non-missing data entry but differ in their imputed values. The M multiple imputed data sets are then analyzed separately using standard methods as if they were complete data sets. As a result of the analysis carried out on the M data sets, the M estimates of each parameter and the related standard errors $[\hat{\theta}^{(m)}; \sqrt{V^{(m)}}]$ are pooled in a single statement using Rubins' rules

(Rubin, 1976). Specifically, denoting the overall estimate with $\bar{\theta}$, the mean of parameter estimates taken over the M data sets is

$$\bar{\theta} = M^{-1} \sum_{m=1}^{M} \hat{\theta}^{m}. \tag{1}$$

The total uncertainty is a weighted sum of the average *within imputation variance* (W) and the *between-imputation variance* (B): $T = W + (1 + M^{-1})B$. The *within variance* ($W = \sum_{m=1}^{M} V^{m}$) is considered the variance that we would observe if there were not missing values in the data set, while the *between variance* ($B = (M-1)^{-1} \sum_{m=1}^{M} (\hat{\theta}^{(m)} - \bar{\theta})^2$) accounts for the uncertainty on the true value of θ due to multiple imputation.

Multiple imputation methods for dealing with multi-item scales are in general borrowed from multiple imputation procedures developed for categorical data. These approaches for imputing multivariate categorical data include joint and conditional modelling methods (Van Buuren and Oudshoorn, 2011; Wu, Jia, and Enders, 2015): e.g. MI, MICE, SRI, MILCA (Raghunathan et al., 2001; Sulis and Porcu, 2008; Vermunt et al., 2008; Van Buuren and Oudshoorn, 2011; Sulis, 2013.) Sulis (2013) carried out a small simulation study in IRT framework to provide a first insight on MICE and MILCA accuracy in the estimation of item parameters. Results highlight that the two procedures provide similar results under ignorable missing data mechanisms when the rate of missing data ranges from 5% up to 30%. Finch (2010) investigates the accuracy of imputation methods for imputing missing categorical data using an ordinal logistic regression model. He compares SRI with (i) MI[1] (a well established multiple imputation method for continuous variables based on the assumption that variables have a multivariate normal probability distribution) and (ii) an ad hoc multiple imputation method for missing categorical data based on the Multinomial distribution (MIC). Results suggest that MI and SRI are competing approaches under ignorable missing data generating processes and both are superior to MIC in reproducing the parameters of the ordered logistic model. SRI displays a slight greater bias in the estimation of parameters than MI under MCAR and similar bias under MAR. In both cases SRI provides lower standard errors than MI. Finch's (Finch, 2008) study designed to assess the accuracy of a wide range of missing data handling methods on the estimates of a three parameter IRT model for dichotomous items (Birnbaum, 1968) with several missing data generating processes concludes that no one method stands as superior

[1] MI was adapted to deal with ordered items by rounding non integer values to conform to the features of the data.

in all cases (with regard to the estimation of all parameters), although MI is frequently associated with slightly lower estimation bias, particularly under MAR condition. Moreover, among the procedures under comparisons, MI produces estimates of the proportion of correct cases which are the closest to the real values. MI appears as preferable to other approaches also in a further validation design study (Finch, 2011) which focuses on the assessment of the impact of missing data handling methods on the detection of nonuniform differential item functioning. Researchers highlight that the performance of multiple imputation methods decreases when model for normal data are fitted to ordinal data and that MI can perform differently with different type of items or IRT models (Ake, 2005; Finch, 2011).

Hereafter, we restrict the attention in the simulation study to the comparison of MILCA with other model-based multiple imputation procedures which displayed some features in terms of effectiveness with categorical items (or which have not been yet validated in IRT framework for ordered data). MI has been selected since it is recommended as championed approach in many previous studies (Finch, 2008, 2010, 2011). SRI has been selected since (i) it proves good performances in dealing with categorical data under medium-low rate of missingness and MAR (Sulis and Porcu, 2008), (ii) its potential in IRT models for ordered data has not been examined in previous studies and (iii) it is a competitor of MI in the estimation of parameters in the logistic regression framework (Finch, 2010). MICE has been selected since in a previous explorative study carried out by Sulis (2013) it shows performances similar to MILCA under ignorable missing data mechanisms (Sulis, 2013) (none of the two methods emerge as superior under all conditions) but the two methods have not been compared under non ignorable missing data generating process. Indeed, we also considered in the simulation study the Relative Mean Substitution (RMS), a deterministic no model-based imputation method. It has been selected for the purpose of making comparison considering its ease of implementation for a non practitioner and because it has been specifically designed for dealing with Likert-type scales. Moreover, simulation studies in the IRT framework (Bernaards and Sijtsma, 1999) have detected that it is superior to random imputation, mean imputation and pairwise deletion techniques (Schafer and Graham, 2002; Huisman, 1999). In the following section we pursue two aims: (i) to assess the performance of MILCA (Sulis, 2013) in the IRT framework for ordered data and to ascertain its potential under several conditions, such as when the ignorability assumption does not hold; (ii) to provide recommendations on which imputation method to use under the possible scenarios described in the simulation study. An advantage of the imputation methods proposed is that they can be adopted even with large scales of items.

4. Multiple Imputation by Latent Class Analysis

Multiple Imputation by Latent Class Analysis (MILCA) (Sulis, 2013) is a model-based multiple imputation technique which relies on Latent Class Analysis (LCA) to generate plausible values for missing observations. LCA has great potential in dealing with missingness since units clustered in the same class share the same expected values for providing responses in the categories of the items composing the scale. Vermunt et al. (2008) show that LCA is a sound modeling approach which overcomes many limits of imputation procedures applied in IRT for the following reasons: i) it considers responses to items as draws from Multinomial distributions; ii) it can detect complex higher order interaction among items; iii) it can be applied to scales with any pattern of missing values and any number of items; iv) it provides reliable estimates of the parameters even under severe rates of missingness; v) it allows us to deal with the uncertainty of parameter values by drawing multiple plausible values. For a comprehensive discussion on the potential of LCA in imputation contexts see Vermunt *et al.* (2008). In the following, we briefly introduce LCA analysis and then we discuss how the MILCA procedure works using the poLCA function implemented in R to carry out LCA (Linzer and Lewis, 2011).

4.1 Latent Class Analysis

LCA is a multivariate statistical analysis technique which allows us to identify a number of categorical unordered latent classes from a multiway table that contains the cross classification of responses to several items. Thus, respondents are classified into R $(r = 1, \ldots, R)$ latent classes on the basis of their joint response pattern to a set of J $(j = 1, \ldots, J)$ items. Specifically, each latent class is identified by two sets of parameters: the *latent class membership probability*, namely p_r, which denotes the proportion of respondents classified in class *r*, and the *item response probability conditional upon the latent class membership*, namely π_{rjk} for $k = 1, \ldots, K$, which defines the probability that respondents in class r select category k of item j. Let us denote with y_{ijk} the indicator variable which takes value 1 if respondent i $(i = 1, \ldots, n)$ selects category k $(k = 1, \ldots, K_j$ the categories) of item j, the joint probability density function of $\boldsymbol{y}_i$ is specified as function of π_{rjk} and p_r

$$P(\boldsymbol{y_i}|\boldsymbol{p}\,,\boldsymbol{\pi}) = \sum_{r=1}^{R} p_r \prod_{j=1}^{J}\prod_{k=1}^{K} (\pi_{rjk})^{y_{ijk}}; \qquad (2)$$

individuals are then classified into classes on the basis of their posterior class membership probabilities (using Bayes' rule).

Table 1. Example of data matrix affected by missingness

unit	items							
♯	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
1	2	2	3	2	2	·	3	·
2	3	3	3	3	·	2	·	3
3	3	3	3	·	3	2	4	3
4	4	4	·	·	3	4	3	·
5	1	2	2	2	1	·	2	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

The poLCA function requires to substitute labels of the categories with subsequent numbers

$$\hat{P}_{(r_i|\boldsymbol{y}_i)} = \frac{\hat{p}_r f(\boldsymbol{y}_i;\hat{\pi}_r)}{\sum_{r=1}^{R} f(\boldsymbol{y}_i;\hat{\pi}_r)} \qquad r = 1, \ldots, R;$$

which for each unit i is a function of the observed response pattern ($\boldsymbol{y}_i$) and of the parameter estimates $\hat{\pi}_{rjk}$ and $\hat{p}_r$.

The poLCA package (Linzer and Lewis, 2011) maximizes the Log-likelihood function with respect to $\hat{\pi}_{rjk}$ and $\hat{p}_r$ using an Expectation-Maximization algorithm.

4.2 How the MILCA Procedure Works

The MILCA procedure uses the poLCA function implemented in R language to apply LCA to a data set of categorical items (Y) with K categories of responses. The missing response in any of the items is considered as a possible response category and it is replaced with a label, namely '$K+1$'. In the following we use the data matrix depicted in Table 1 as an example to illustrate how the procedure works step by step:

1. Missing values are recoded in the category '$K+1$' (i.e. category 5 in Table 2);
2. LCA is applied to the data matrix described in Table 2;
3. The main results provided by LCA are the estimates of the vector of *latent class membership probabilities* ($\hat{p}_r$) and the *item response probabilities* conditional upon the class membership ($\hat{\pi}_{jrk}$) (e.g., for a model with three classes, the parameters related to the data matrix described in Table 2 are listed in Table 3);
4. On the basis of both the observed vector of responses (y_i) and the parameter estimates ($\hat{\pi}_{rjk}$ and $\hat{p}_r$), the posterior class membership probabilities $\hat{P}(y_i|r)$ of each unit i are calculated using Bayes' rule (see Table 4);

5. Units (individuals) are classified in one of the R classes on the basis of their modal posterior probability (see last column of Table 4);
6. For each unit i a missing value in item j (for $j = 1, \ldots, J$) is replaced by generating a random draw from a Multinomial distribution with the vector of parameters equal to the estimated vector of *item response probabilities* of the class where the unit has been classified in Step 5: $\hat{\boldsymbol{\pi}}_{jr}(\hat{\pi}_{jr1}, \ldots, \hat{\pi}_{jr(K+1)})$ (see Table 3); for instance for respondent $i = 1$, classified in latent class $r = 2$, the missing observation in item y_8 is imputed by generating M valid random values from a Multinomial distribution with the vector of probabilities equal to the estimated vector of *item response probabilities* for item y_8 in class $r = 2$, namely: Multinomial(0.158, 0.094, 0.264, 0.413, 0.071) (see Table 3). The generated value (category) is valid if it is different from the missing category, namely '$K + 1$';
7. If a random generated value is equal to the code of the missing category $(K + 1)$ (e.g., identified by value 5 in Table 2), the value is not considered as plausible for imputation purposes and it is rejected. The procedure is iterated until a new value different from the code for missing is generated. Let us suppose that the $M = 6$ random draws for imputing a missing value in an item with four category are equal to '3', '3', '4', '5=missing', '4' and '2'. The value '5' is not considered a valid draw because it corresponds to the code of the missing category. Thus a new value is drawn. The procedure is iterated until M valid draws are generated for each missing value in the data.
8. The M values are used to generate M imputed data sets $(Y^1, \ldots, Y^M)$ that are identical for the non-missing data entry but differ in their imputed values
9. The M datasets are analyzed using MIA.

MILCA explicitly takes into account the information on missingness in defining the latent class parameters by considering the empty observation as a response category. In standard Latent Class Analysis, models are selected according to the parsimony criterion by minimising the Akaike Information Criterium (AIC) or the Bayesian Information Criterium (BIC). The latter is preferred to the former when the latent class membership probability is not specified as a function of covariates, as it is in the MILCA procedure (Linzer and Lewis, 2011). The number of LCs in MILCA is selected by applying the LCA models with a different number of LCs to the data matrix in which missing values are recorded in the category $(K + 1)$ (see Step 2 of the MILCA procedure and Table 2) and selecting the model which provides the lowest value of the indexes (Nylund, Asparouhov, and Bengt, 2007). In

Table 2. Missing values recoded in the first step of MILCA

unit	items							
$\sharp$	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
1	2	2	3	2	2	5	3	5
2	3	3	3	3	5	2	5	3
3	3	3	3	5	3	2	4	3
4	4	4	5	5	3	4	3	5
5	1	2	2	2	1	5	2	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

The poLCA function requires to substitute labels of the categories with subsequent numbers

Table 3. LCA estimates – example from fitting a 3 latent classes model

Classes	$\hat{\pi}_{rj1}$	$\hat{\pi}_{rj2}$	$\hat{\pi}_{rj3}$	$\hat{\pi}_{rj4}$	$\hat{\pi}_{rj5}$
		y_1			
$r = 1$	0.0029	0.0265	0.3822	0.5435	0.0450
$r = 2$	0.4941	0.3623	0.0660	0.0149	0.0627
$r = 3$	0.0331	0.2870	0.5512	0.0865	0.0423
		y_2			
$r = 1$	0.0037	0.0000	0.1292	0.8144	0.0526
$r = 2$	0.2090	0.3366	0.2830	0.0982	0.0731
$r = 3$	0.0100	0.1101	0.5724	0.2617	0.0457
⋮	⋮	⋮	⋮	⋮	⋮
		y_8			
$r = 1$	0.0031	0.0000	0.0475	0.8919	0.0574
$r = 2$	0.1580	0.0945	0.2637	0.4132	0.0706
$r = 3$	0.0070	0.0308	0.2626	0.6508	0.0487
		Estimated class population shares $\hat{p}_r$			
$r_1 = 0.314$	$r_2 = 0.228$	$r_3 = 0.457$			
		Predicted class memberships by modal posterior			
$r_1 = 0.316$	$r_2 = 0.227$	$r_3 = 0.457$			

Table 4. Posterior membership probabilities and modal assignment: $\hat{P}(\boldsymbol{y}_i|r)$

$\sharp$	$\hat{P}(r = 1)$	$\hat{P}(r = 2)$	$\hat{P}(r = 3)$	Modal class
1	0.00	0.96	0.00	$r = 2$
2	0.00	0.00	1.00	$r = 3$
3	0.00	0.00	1.00	$r = 3$
4	0.99	0.00	0.01	$r = 1$
5	0.00	1.00	0.00	$r = 2$
⋮	⋮	⋮	⋮	

one of the simulation studies a sensitivity analysis is carried out to assess the influence of the criterium adopted on the final result.

For the purposes of comparisons, in the the following we will present a short description of the other imputation methods that are used in the simulation studies.

4.3 Multivariate Imputation

The Multivariate Imputation (MI) model is based on the assumption that the probability model underlying a set of variables (the $\boldsymbol{Y}$ matrix) is multivariate normal. The method works by iterating two steps (Shafer, 1997; Shafer and Graham, 2002; Wu, Jia, and Enders, 2015):

1. In the Posterior Step, M random values of the parameters $\boldsymbol{\theta}[\theta_1, \ldots, \theta_p]$ of the multivariate normal distributions are drawn from their posterior distribution, namely $\boldsymbol{\theta} \sim P(\boldsymbol{\theta}|Y_{obs}, Y_{miss})$
2. In the Prediction Step M missing values are generated as random draws from the predictive distribution of Y_{miss}, namely $Y_{miss} \sim P(Y_{miss}|Y_{obs}, \boldsymbol{\theta})$
3. The two steps described at point 1 and 2 are iterated until the posterior distribution of the parameters is stabilized
4. The M predicted values are used to generate M imputed data sets $(Y^1, \ldots, Y^M)$ that are identical for the non-missing data entry but differ in their imputed values
5. The non integer values are rounded to be adapted to the scale of the items
6. The M datasets are analyzed using MIA.

4.4 Simple Imputation Methods: Relative Mean Substitution

The Relative Mean Substitution (RMS) (Raaijmakers, 1999; Finch, 2008) replaces a missing value y_{ij} by weighting the mean of the item calculated on non missing responses (Total Mean Substitution – $TMS_{.j}$) with the ratio between the intra-individual mean of the respondent i for all non missing items (Valid Mean Substitution – $VMS_{i.}$) and the sample mean of the other respondents (excluding respondent i) for the same items (GMS^{-i})

$$RMS(y_{ij}) = \frac{VMS_{i.}}{GMS^{-i}} TMS_{.j}. \tag{3}$$

The ratio indicates the relative position of the mean of the responses provided by individual i to the non missing items with respect to the overall mean for all other respondents for the same items. The weight is larger than 1 for respondents with scores higher than the average.

This method has the advantage of being easily computed whenever missing scores are observed on several items; it handles Likert-scales as metrical, thus it assigns subsequent numbers to adjacent categories.

4.5 Sequential Regression Imputation Methods

Imputation methods based on fully conditional approaches give enormous flexibility in predicting missing values in large datasets whenever several variables are affected by missingness (Raghunathan et al., 2001; Little and Rubin, 2002; Sulis and Porcu, 2008; Van Buuren and Oudshoorn, 2011). This class of methods solves the multivariate imputation model for a matrix Y of items affected by missingness using a variable by variable imputation approach (Van Buuren and Oudshoorn, 2011). The approach consists in specifying a set of sequential and univariate conditional densities, where plausible values for each item y_j are generated conditional upon the remaining items (denoted as $\boldsymbol{Y}-j$). The system of equations is sequentially iterated and at each iteration new plausible values are drawn and the imputed values are updated. Several adaptations of sequential multiple imputation have been advanced in the literature (Raghunathan et al., 2001; Sulis and Porcu, 2008; Van Buuren and Oudshoorn, 2011). Here, we consider MICE and SRI. Both procedures have been implemented with specific functions in R (Sulis and Porcu, 2008; Van Buuren and Oudshoorn, 2011).

4.5.1 Multiple Imputation by Chained Equation

The Multiple Imputation by Chained Equation (MICE) algorithm implemented in R (Van Buuren and Oudshoorn, 2011; Wu, Jia, and Enders, 2015) is a fully conditional approach which consists of two steps that are sequentially iterated for each of the J variables affected by missingness. At each iteration t the algorithm works as follows:

1. In the Posterior Step the parameter vector $\boldsymbol{\theta}_j^{(t)}$ of the probability distribution of the imputation parameters of item $y_j^t(y_{j.obs}, y_{j.imp})$ are generated conditional upon the values of the other items

   $\theta_j^{(t)} \sim P(\theta_j^{(t)}|y_{j.obs}, y_1^t, ...y_{j-1}^t, y_{j+1}^{t-1}, ...y_J^{t-1})$.

2. In the Prediction Step missing values for item $y_{j.imp}^{(t)}$ are replaced by draws from their conditional distribution

   $y_{j.imp}^{(t)} \sim P(y_j|y_{j.obs}, y_1^t, ...y_{j-1}^t, y_{j+1}^{t-1}, ...y_J^{t-1}, \theta_j^{(t)})$.

3. Once convergence of the parameters $\boldsymbol{\theta}$ is reached, M plausible values are generated for each missing value in Y.
4. The M values are used to generate M imputed data sets $(\boldsymbol{Y}^1, \ldots, \boldsymbol{Y}^M)$ that are identical for the non-missing data entry but differ in their imputed values.
5. The M datasets are analyzed using MIA.

The probabilistic model is selected according to the scale of variables. For categorical variables the package invokes the `polyreg` function which specifies a multinomial logit model (Agresti, 2002). More details on how MICE works are provided in Van Buuren and Oudshoorn (2011).

4.5.2 Multiple Imputation by Stochastic Regression

The Multiple Imputation by Stochastic Regression (MISR) procedure (Sulis and Porcu, 2008) is a stochastic imputation procedure (implemented in R language) which works as follows:

1. For each unit i the procedure stars building up the marginal distribution of responses in each of the K response categories $(\pi_{i1}, \pi_{i2}, \ldots, \pi_{iK})$. Missing values of unit i in any item are replaced by drawing M values from a Multinomial distribution with parameters set equal to the proportion of responses observed in each response category.
2. The M random draws generated for each missing value are used to create M data sets $(\boldsymbol{Y}^1, \ldots, \boldsymbol{Y}^M)$ that are identical for the non-missing data entry but differ in their imputed values.
3. Next, in each of the M data sets, y_j (for $j = 1, \ldots, J$) is modeled conditional upon the remaining items, namely $\boldsymbol{Y}(-j)$, using an ordinal logistic model. For each $y_{ij.imp}$ the predicted vector of conditional probability (i.e. $\hat{\pi}_{ij1}(\boldsymbol{Y}(-j)), \ldots, \hat{\pi}_{ijK}(\boldsymbol{Y}(-j))$) is used to generate a random draw from a Multinomial distribution.
4. The M values are replaced in M data sets $(\boldsymbol{Y}^1, \ldots, \boldsymbol{Y}^M)$ that are identical for the non-missing data entry but differ in their imputed values.
5. The M datasets are analyzed using MIA

5. A Simulation Study to Validate the Accuracy of MILCA for IRT Models

5.1 IRT Models

The accuracy of the compared procedures in the estimation of item parameters was assessed using the most popular IRT model for ordinal items, the Graded Response Model (Samejima, 1969). The model specifies the logit of the cumulative probability that unit i selects a category not lower than k $\{\gamma_{ijk}\}$ of item y_{ij} in terms of item and person parameters

$$logit(\gamma_{ijk}) = \lambda_j(\eta_i - \beta_{jk}), \tag{4}$$

where, β_{jk} is the category-threshold parameter (of category k with the lower), λ_j is the discrimination parameter and η_i is a person parameter. Parameter η_i is considered a random effect with probability distribution $\mathcal{N}(0,1)$.

The total number of parameters is $(K-1) \times (J)$ category-threshold parameters and J discrimination parameters. Function Grm in the ltm package (Rizopoulos, 2006; Linzer and Lewis, 2011) from R uses Gauss-Hermite quadrature to approximate the marginal likelihood and a Newton-Raphson algorithm to maximize it.

5.2 Measuring the Accuracy in Estimation of Item and Person Parameters

The accuracy in estimation was evaluated by calculating two measures of accuracy for each parameter (item-category and discrimination) of the Graded Response Model; each measure considers the extent to which the imputation procedures preserve the true value of the parameters (estimated on the benchmark data) as well as the efficiency of the estimates.

The classic Mean Squared Error measure of a parameter θ

$$MSE(\hat{\theta}) = (\hat{\theta} - \theta)^2 + Var(\hat{\theta}) \qquad \forall\, \hat{\theta}, \theta \neq 0.$$

evaluates the accuracy of the estimates making a tradeoff between bias and efficiency which depends on absolute differences across parameters.

A second measure was introduced by the authors to consider the extent to which the estimates of the parameters differ from the true values in relative terms, at the same time balancing for efficiency. The Relative Accuracy Index (RAI) was defined as

$$RAI(\hat{\theta}) = (\frac{\hat{\theta}}{\theta} - 1)^2 + Var(\hat{\theta}) \qquad \forall\, \hat{\theta}, \theta \neq 0.$$

To facilitate an assessment of the overall accuracy of the imputation methods in terms of MSE and RAI, both indexes were summarized by taking the sum over the threshold and discrimination parameters. Specifically, the following overall measures of accuracy of item-threshold and discrimination parameters were defined for each model:

1. The Model Overall Mean Squared Error of the threshold parameters

    - threshold parameters

$$MOMSE_{\beta} = \sum_{j} \sum_{k} MSE(\beta_{jk}) \tag{5}$$

    - discrimination parameters

$$MOMSE_{\lambda} = \sum_{j} MSE(\lambda_{j}) \tag{6}$$

    - a pooled measure of both

$$MOMSE_{\beta,\lambda} = MOMSE_{\beta} + MOMSE_{\lambda} \tag{7}$$

2. The Model Overall Relative Accuracy Index of

    - threshold parameters

$$MORAI_{\beta} = \sum_{j} \sum_{k} RAI(\beta_{jk}) \tag{8}$$

    - discrimination parameters

$$MORAI_{\lambda} = \sum_{j} RAI(\lambda_{j}) \tag{9}$$

    - a pooled measure of both

$$MORAI_{\beta,\lambda} = MORAI_{\beta} + MORAI_{\lambda} \tag{10}$$

For all the indexes, the higher their values the worse the overall estimation of accuracy for the related imputation procedure. We suggest to use the overall $MOMSE_{\beta,\lambda}$ and $MORAI_{\beta,\lambda}$ indexes only as a first screening tools to assess the overall size of the departure. Given that threshold and discrimination parameters are presumably on different scales, it is recommended to look at the single components.

5.3 Simulation Design

The MILCA procedure was validated developing two simulation studies on two complete data sets (without missing data) : (i) a data set from a survey on students' evaluation of teaching in a university containing a scale addressed to measure students' perceived quality and (ii) a data set from the PIRLS survey 2011 containing a scale addressed to measure students' attitude towards reading (Mullis et al., 2012). The article aims to assess MILCA accuracy and examines the extent to which the choice of the imputation procedure influences the estimates of item parameters under different missing data mechanisms and under two different scenarios in terms of number of items of the measurement instrument and sample size. Missing values were generated in both complete datasets (used as banchmark) deleting observations from items according to three different missing data generating processes: MCAR, MAR and NMAR. The three missing data mechanisms were simulated using functions miss.CAR, miss.AR, miss.NAR written by the authors in R language.

In both simulation studies (Simulation 1 and 2), missing values were generated in the complete data set according to the three missing generating processes (MCAR, MAR, MNAR) and six rates of missingness (π=5%, 10%, 15%, 20%, 25%, and 30%). As a result for each complete data set 18 data sets affected by missingness were generated (6 for each missing generating process) and imputed with the MILCA procedure, and, for comparative purposes, the results were compared with the other four imputation methods, namely MI, RMS, MICE and MISR.

Taking a situation of MCAR an observation was set as missing if the result of random draw from a Bernoulli with parameter π(π=5%, 10%, 15%, 20%, 25%, and 30%) was 1.

Under a MAR condition the probability of setting an observation as missing depends on certain observed covariates (see function miss.CAR). Under a MAR condition a unit i in the matrix was set as missing if the result of a random draw from a Bernoulli with parameter estimated as function of individual predictors ($\hat{\pi}_i(\boldsymbol{x})$) was 1 (see function miss.MAR), where

$$\pi_i(\boldsymbol{x}) = \frac{\exp(\boldsymbol{\beta}'\boldsymbol{x}_i)}{1 + \exp(\boldsymbol{\beta}'\boldsymbol{x}_i)}. \tag{11}$$

The MNAR scenario was simulated by fixing the probability (π_i) an observation being set as missing according to the intensity of the individual value on the latent trait. Specifically, individuals are clustered in four classes on the basis of the quartiles of the distribution of an individual's sum of scores. Different degrees of probability of skipping an item were applied to individuals' belonging to each of the four quartiles (see function miss.NAR).

Simulation 1: The data set on students' evaluation of teaching includes 8 items ($y_1 - y_8$) addressed to measure teaching quality in students' perception, one to student's attendance at lectures (A) and one to student's interest toward the topic taught (I) (see Table 5). The complete data set contains 1737 observations.

Table 5. Item considered for the application

Item	Contents
y_1	*Lecturer motivates students*
y_2	*Lecturer highlights topics*
y_3	*Lecturer answers questions during the class*
y_4	*Lecturer clarifies goals of the course*
y_5	*Lecturer clearly explains topics*
y_6	*Lecturer suggests how to study*
y_7	*Lecturer gives classes on schedule*
y_8	*Global satisfaction*
$x_1 = A$	*Student's attendance at classes*
$x_2 = I$	*Student's interest toward the topics*

All items are measured on a four-category Likert scale: *Definitely No*, *More No than Yes*, *More Yes than No*, *Definitely Yes*. The probability of skipping an item in the application is assumed to depend on two students' covariates. Specifically, *Students' attendance at classes* ($1 =$ *Always*; $4 =$ *Very rarely*) and *Students' interest toward the topic* ($1 =$ *Definitely No*; $4 =$ *Definitely Yes*). In the complete data set, the cross-classification of units according to these two covariates provides 16 groups of students. Values have been set MAR by attaching to each of the 16 groups a different degree of probability (π) of skipping an item using equation 11. The β parameter vector estimates was defined attaching the lowest probability π_i of skipping an item to students who say they are interested (*Definitely Yes*) in the topic and who have *Always* attended the classes; the highest $\pi_i(x)$ is attached to students who say they are *Definitely No* interested and who have attended classes *Rarely*. Values have been set MNAR by attaching the lowest probability to skip an item to the quartile with the lowest level of the latent trait, the highest to the ones with the highest level.

Simulation 2: The data set includes 13 items related to student's attitude to reading ($y_1 - y_{13}$), one item related to the number of books available at home ($Books$) and gender (G: 1 Female, 0 Male) (see Table 6). The data set contains 3608 observations.

All items are measured on a three-category Likert scale: *Agree a lot*, *Agree a little*, *Disagree*. Under the MAR condition the probability of skipping an item is considered to depend on the number of books at home *Books* (from $1 =$ *more than 200* to $5 =$ *0-10*) and *Gender* ($1 =$ *Female*; $0 =$ *Male*). The cross-classification of units according to these two covariates provides 10 groups of students; each with a different degree of probability ($\pi(x)$) of skipping an item. The β parameter vector estimates was defined attaching the lowest probability π_i of skipping an item to female students with more

Table 6. Item considered for the application

Item	dir.	Contents
y_1	$-$	*I read only...*
y_2	$+$	*I like talking ...*
y_3	$+$	*I would be happy ...*
y_4	$-$	*I think reading ...*
y_5	$+$	*I would like to ...*
y_6	$+$	*I usually do ...*
y_7	$+$	*Reading is easy ...*
y_8	$-$	*Reading is harder ...*
y_9	$+$	*If a book is ...*
y_{10}	$-$	*I have trouble ...*
y_{11}	$+$	*My teacher tells ...*
y_{12}	$-$	*Reading is harder ...*
y_{13}	$+$	*I like to read ...*
$x_1 = BOOKS$		*number of books at home*
$x_1 = GENDER$		

Items belong to sections G4, R7, R8 and R9 of PIRLS 2011 Student Questionnaire
Source: PIRLS 2011 User Guide for the International Database. Copyright ©2013 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center,Lynch School of Education, Boston College.
Negative items have been reversed
Response items *Agree a lot*=1 *Agree a little*=2 *Disagree*=3

than 200 books at home. Values have been set MNAR by attaching the the lowest probability to skip an item to the quartile with the lowest level of the latent trait, the highest to the one with the highest level.

5.4 Results

Results of the estimates of the item-threshold and item-discrimination parameters for all the fitted models are listed in the supplementary online materials (Tables S1–S5 Simulation 1 and S6–S8 Simulation 2). The Tables displayed the ratio between the estimates of the parameters obtained by using a specific imputation procedure (the procedures are listed in the columns) and the estimates observed on the benchmark data sets.

The eighteen datasets generated in each simulation study were imputed using MI, MICE, MISR, RMS and MILCA. The accuracy of the methods was then evaluated using the MOMSE and MORAI measures to overcome the difficulty of highlighting the accuracy of the five imputation procedures in each of the 18 data sets (for each simulation study) by looking at the single estimates of the item-category and discrimination parameters (see Tables 8 and 9).

5.4.1 Simulation 1

The MILCA function was applied by defining 4 different numbers of LCs (ranging from 3 to 6) for each of the 18 datasets. We compare the results obtained by selecting the number of LCs according to the BIC or AIC criteria to identify which criterion identifies the imputation model with the best parameter prediction accuracy. Furthermore, we also compared the item parameter prediction accuracy provided by data sets imputed using the MILCA procedure with a different number of LCs (ranging from 3 to 6) to assess how the choice of the number of LCs affects the accuracy in estimation (Vermunt et al., 2008; Sulis, 2013). Tables 8 and 9 list the MOMSE and MORAI indexes for all missing imputation procedures considered in this study, under the three missing data generating processes and the six different percentages of missingness.

In Simulation 1, the analysis pursues two main tasks: (i) to highlight which goodness of fit criterion is recommended in order to select the number of LCs under the three missing data generating processes and the different rates of missingness, (ii) to compare MILCA with the other imputation procedures in order to assess under which conditions it has a greater effectiveness.

Missing Completely at Random: Table 7 shows that for MCAR observations the BIC index would recommend selecting 5 LCs when the percentage of missingness is medium-low (5% or 10%), 4 LCs when it is medium-high (15 and 25%), and 3 LCs when it is severe (about 30%). Following the AIC criterion, 6 LCs are always recommended. The values of the MOMSE indexes for the related models are listed in Table 8.

Looking at the $MOMSE_{\beta,\lambda}$ index it emerges that until the rate of missingness is approximately 20% the three procedures may be considered equivalent; however, MILCA and MICE prove to be still good when the rate of missingness increases. The RMS and MI seem to be a competitive alternative only for a low rate of missingness (10%). The values of the MOMSE index calculated on datasets imputed using different numbers of LCs (see Table 8) suggest that the MILCA procedure is weakly influenced by the choice of LCs under the MCAR condition. By selecting the number of LCs recommended under the BIC criterion it emerges that in 5 out 6 scenarios (rates of missingness) MILCA provides values of the $MOMSE_{\beta,\lambda}$ index almost equal (4 out of 5 times) or better (1 out of 5 times) than MICE. The selection of the number of LCs on the basis of the AIC criterion would lead to a slightly weaker result in 4 out of 6 scenarios and to a slightly better one in 2 out 6.

The closeness of the estimates using MILCA and MICE is confirmed also by the comparison of the accuracy of the estimates in relative terms us-

Table 7. Goodness of fit measures: Models with different number of LCs

			5%	10%	15%	20%	25%	30%
					MCAR			
MILCA	3LC	AIC(3):	31689.54	34011.16	35341.59	36241.71	36611.9	36536.37
		BIC(3):	32224.55	34546.18	35876.60	36776.72	37146.91	**37071.38**
		$G^2(3)$:	8197.89	9580.86	10378.70	10922.18	11110.87	10913.69
	4LC	AIC(4):	31326.47	33713.29	35051.47	36011.45	36429.80	36382.55
		BIC(4):	32041.64	34428.46	**35766.65**	**36726.62**	**37144.97**	37097.73
		$G^2(4)$:	7768.82	9216.989	10022.58	10625.92	10862.77	10693.88
	5LC	AIC(5):	31098.87	33516.45	34880.25	35862.05	36318.29	36293.77
		BIC(5):	**31994.20**	**34411.78**	35775.58	36757.39	37213.62	37189.10
		$G^2(5)$:	7475.22	8954.15	9785.36	10410.52	10685.26	10539.09
	6LC	AIC(6):	31007.93	33428.66	34809.18	35810.89	36270.61	36270.63
		BIC(6):	32083.42	34504.15	35884.67	36886.38	37346.1	37346.12
		$G^2(6)$:	7318.28	8800.36	9648.29	10293.36	10571.58	10449.95
					MAR			
MILCA	3LC	AIC(3):	31371.68	33600.56	35238.48	35777.42	36168.94	36214.58
		BIC(3):	31906.7	34135.58	35773.49	36312.43	36703.96	36749.59
		$G^2(3)$:	8061.024	9547.661	10479.28	10768.06	10991.7	10982.63
	4LC	AIC(4):	31044.48	33293.08	34973.92	35538.18	35983.56	35891.20
		BIC(4):	31759.65	34008.25	**35689.1**	**36253.35**	**36698.73**	**36606.37**
		$G^2(4)$:	7667.821	9174.176	10148.72	10462.82	10740.32	10593.24
	5LC	AIC(5):	30815.19	33075.80	34811.32	35384.14	35756.24	35749.61
		BIC(5):	**31710.52**	**33971.13**	35706.65	36279.47	36651.57	36644.94
		$G^2(5)$:	7372.529	8890.89	9920.11	10242.78	10447.00	10385.65
	6LC	AIC(6):	30720.01	32983.18	34737.72	35239.43	35617.34	35637.54
		BIC(6):	31795.50	34058.67	35813.21	36314.92	36692.83	36713.03
		$G^2(6)$:	7211.35	8732.278	9780.52	10032.08	10242.09	10207.58
					MNAR			
MILCA	3LC	AIC(3):	31584.32	33915.45	35135.73	35775.64	36159.53	35930.92
		BIC(3):	32119.33	34450.47	35670.75	36310.65	36694.54	36465.94
		$G^2(3)$:	7914.905	9156.421	9997.358	10403.05	10649.93	10433.19
	4LC	AIC(4):	31241.75	33573.45	34791.43	35439.81	35832.7	35610.96
		BIC(4):	**31956.92**	**34288.63**	**35506.61**	**36154.98**	36547.87	36326.13
		$G^2(4)$:	7506.339	8748.421	9587.059	10001.22	10257.1	10047.22
	5LC	AIC(5):	31107.13	33456.81	34647.77	35328.46	35741.65	35530.96
		BIC(5):	32002.46	34352.14	35543.1	36223.8	36636.98	36426.29
		$G^2(5)$:	7305.718	8565.775	9377.393	9823.879	10100.05	9901.226
	6LC	AIC(6):	30922.75	33292.2	34555.96	35250.11	35640.46	35430.66
		BIC(6):	31998.24	34367.69	35631.45	36325.60	**36715.95**	**36506.15**
		$G^2(6)$:	7055.341	8335.169	9219.585	9679.527	9932.862	9734.925

*Classes selected with the BIC criterion

Table 8. **M**odel **O**verall **M**ean **S**quare **E**rror of GRM parameters

		5%	10%	15%	20%	25%	30%
MCAR							
$MOMSE_{\beta}$ (a)							
MI		1.078	1.133	1.246	1.482	1.515	1.728
MICE		1.065	1.059	1.061	1.045	1.072	1.102
MISR		1.065	0.981	0.980	0.938	1.000	1.131
RMS		1.353	1.681	2.240	2.546	3.495	4.796
MILCA	3LC	1.004	1.032	1.036	1.036	0.978	0.975
	4LC	1.058	1.067	1.003	1.025	1.079	1.089
	5LC	1.068	1.078	1.126	1.090	1.105	1.208
	6LC	1.075	1.096	1.151	1.121	1.132	1.208
$MOMSE_{\lambda}$ (b)							
MI		0.125	0.200	0.264	0.438	0.668	0.624
MICE		0.118	0.175	0.167	0.234	0.312	0.406
MISR		0.128	0.183	0.223	0.391	0.550	0.685
RMS		0.160	0.311	0.623	1.126	2.092	3.395
MILCA	3LC	0.137	0.256	0.361	0.401	0.553	0.479
	4LC	0.122	0.149	0.226	0.269	0.349	0.326
	5LC	0.123	0.164	0.190	0.325	0.326	0.283
	6LC	0.118	0.135	0.153	0.231	0.305	0.237
$MOMSE_{\beta,\lambda}$ (c)							
MI		1.203	1.333	1.510	1.919	2.183	2.352
MICE		1.183	1.234	1.228	1.279	1.383	1.507
MISR		1.193	1.163	1.203	1.329	1.550	1.816
RMS		1.513	1.993	2.863	3.672	5.587	8.191
MILCA	3LC	1.141	1.287	1.397	1.437	1.531	**1.454**
	4LC	1.180	1.216	**1.229**	**1.294**	**1.428**	1.415
	5LC	**1.192***	**1.242**	1.316	1.416	1.431	1.491
	6LC	1.193	1.231	1.304	1.352	1.436	1.445
MAR							
$MOMSE_{\beta}$ (d)							
MI		1.074	1.123	1.491	1.877	1.529	1.475
MICE		1.075	1.074	1.122	1.133	1.051	1.066
MISR		1.051	0.994	0.980	0.983	0.986	1.240
RMS		1.313	1.660	2.147	2.427	3.051	3.637
MILCA	3LC	1.030	1.015	1.257	1.131	1.047	0.918
	4LC	1.074	1.059	1.172	1.277	1.190	1.097
	5LC	1.092	1.076	1.233	1.071	0.947	1.162
	6LC	1.083	1.099	1.217	1.127	0.975	0.961
$MOMSE_{\lambda}$ (e)							
MI		0.126	0.157	0.242	0.274	0.379	0.369
MICE		0.120	0.141	0.176	0.179	0.206	0.169
MISR		0.118	0.143	0.230	0.301	0.330	0.458
RMS		0.157	0.322	0.782	1.101	1.769	2.623
MILCA	3LC	0.136	0.221	0.321	0.369	0.393	0.443
	4LC	0.118	0.128	0.143	0.211	0.175	1.238
	5LC	0.113	0.134	0.167	0.288	0.558	0.842
	6LC	0.119	0.143	0.167	0.271	0.435	0.709
$MOMSE_{\beta,\lambda}$ (f)							
MI		1.199	1.279	1.733	2.151	1.908	1.845
MICE		1.195	1.216	1.298	1.312	1.257	1.235
MISR		1.170	1.137	1.210	1.285	1.316	1.698
RMS		1.470	1.982	2.929	3.528	4.821	6.260
MILCA	3LC	1.165	1.236	1.578	1.500	1.441	1.361
	4LC	1.192	1.187	**1.315**	**1.488**	**1.365**	**2.336**
	5LC	**1.206**	**1.210**	1.400	1.359	1.505	2.004
	6LC	1.203	1.242	1.384	1.398	1.410	1.669

(*continued on next page*)

Table 8. **M**odel **O**verall **M**ean **S**quare **E**rror of GRM parameters (*continued*)

		5%	10%	15%	20%	25%	30%
				MNAR (g)			
				$MOMSE_{\beta}$			
MI		1.046	1.266	1.578	1.670	1.930	2.225
MICE		1.037	1.067	1.126	1.085	1.145	1.146
MISR		1.016	1.019	1.039	0.999	0.960	0.992
RMS		1.248	1.872	2.704	3.416	5.072	8.856
MILCA	3LC	0.988	0.978	0.994	0.983	0.967	0.935
	4LC	1.004	1.002	0.996	1.020	1.006	0.950
	5LC	1.032	1.083	1.111	1.062	1.070	0.932
	6LC	1.042	1.096	1.165	1.145	1.028	0.981
				$MOMSE_{\lambda}$ (h)			
MI		0.126	0.289	0.493	0.749	0.919	1.297
MICE		0.112	0.145	0.181	0.270	0.353	0.392
MISR		0.117	0.193	0.232	0.317	0.460	0.437
RMS		0.134	0.280	0.449	0.631	1.042	1.795
MILCA	3LC	0.134	0.238	0.252	0.361	0.524	0.607
	4LC	0.128	0.172	0.232	0.259	0.317	0.405
	5LC	0.128	0.210	0.228	0.211	0.276	0.272
	6LC	0.115	0.170	0.220	0.264	0.217	0.215
				$MOMSE_{\beta,\lambda}$ (i)			
MI		1.172	1.555	2.072	2.419	2.849	3.523
MICE		1.149	1.212	1.307	1.355	1.497	1.537
MISR		1.133	1.213	1.270	1.315	1.420	1.429
RMS		1.382	2.151	3.153	4.046	6.114	10.651
MILCA	3LC	1.121	1.216	1.247	1.345	1.491	1.543
	4LC	**1.132**	**1.174**	**1.229**	**1.279**	1.322	1.354
	5LC	1.160	1.293	1.338	1.272	1.346	1.204
	6LC	1.157	1.265	1.385	1.409	**1.245**	**1.197**

*Classes selected with the BIC criterion

ing the $MORAI_{\lambda,\beta}$ (see Table 9). However, these comparisons are slightly in advantage of MICE if the number of LCs is chosen according to the BIC criterion.

The trend of the overall $MORAI_{\lambda,\beta}$ does not indicate a clear dominant criterion to fix the number of LCs. However, in relative terms, differences across models with different numbers of latent classes can also be considered not relevant: most of the results differ at the second decimal place. It is interesting to highlight that if the focus of the analysis is on the accuracy of the item-discrimination parameters, the $MOMSE_{\lambda}$ index clearly points out that the MILCA procedure implemented with 6 LCs (according to the AIC criterion) provides more accurate estimates compared with the choice of a different number of classes and with the others imputation procedures considered in the study (see Table 8(b)). This result also holds with comparisons between methods in relative terms, as Table 9(b) shows. This evidence would advice to select the number of LCs (to set in the MILCA function) according to the BIC criterion if the aim is to maximize the overall accuracy in absolute terms, but to follow the AIC criterion whenever there is an interest in maximizing the accuracy of the discrimination parameters.

Table 9. **M**odel **O**verall **R**elative **A**ccuracy **I**ndicator of GRM parameters

	5%	10%	15%	20%	25%	30%
MCAR						
$MORAI_{\beta}$ (a)						
MI	1.099	1.156	1.383	1.368	1.562	1.939
MICE	1.086	1.082	1.110	1.066	1.124	1.171
MISR	1.197	1.115	1.150	1.122	1.366	1.449
RMS	1.381	1.698	2.345	2.393	3.328	4.780
MILCA 3LC	1.026	1.029	1.063	1.073	1.308	1.293
4LC	1.068	1.112	1.068	1.096	1.114	1.225
5LC	1.112	1.152	1.290	1.152	1.143	1.578
6LC	1.087	1.220	1.401	1.237	1.177	1.419
MISR	1.077	0.986	1.021	0.973	1.168	1.236
$MORAI_{\lambda}$ (b)						
MI	0.110	0.131	0.138	0.173	0.258	0.226
MICE	0.113	0.133	0.125	0.144	0.146	0.195
MISR	0.120	0.129	0.130	0.150	0.198	0.213
RMS	0.125	0.179	0.289	0.452	0.778	1.187
MILCA 3LC	0.115	0.140	0.149	0.163	0.174	0.165
4LC	0.113	0.120	0.134	0.134	0.182	0.165
5LC	0.113	0.121	0.141	0.140	0.155	0.161
6LC	0.111	0.115	0.126	0.137	0.155	0.164
$M0RAI_{\beta,\lambda}$ (c)						
MI	1.203	1.333	1.510	1.919	2.183	2.352
MICE	1.200	1.215	1.235	1.211	1.271	1.365
MISR	1.197	1.115	1.150	1.122	1.366	1.449
RMS	1.506	1.876	2.634	2.844	4.107	5.968
MILCA 3LC	1.140	1.169	1.212	1.236	1.483	1.458
4LC	1.181	1.232	1.202	1.230	1.296	1.390
5LC	1.225	1.273	1.431	1.292	1.298	1.739
6LC	1.199	1.335	1.527	1.374	1.332	1.583
MAR						
$MORAI_{\beta}$ (d)						
MI	1.141	1.159	1.310	1.397	1.471	1.733
MICE	1.100	1.089	1.148	1.120	1.055	1.154
MISR	1.079	1.004	1.017	1.129	1.196	1.347
RMS	1.340	1.721	2.196	2.559	3.383	3.903
MILCA 3LC	1.037	1.009	1.374	1.250	1.025	1.065
4LC	1.118	1.068	1.117	1.135	1.245	1.440
5LC	1.164	1.095	1.215	1.093	1.079	1.190
6LC	1.127	1.205	1.413	1.183	0.997	1.035
$MORAI_{\lambda}$ (e)						
MI	0.110	0.123	0.153	0.166	0.192	0.176
MICE	0.114	0.126	0.141	0.150	0.172	0.146
MISR	0.110	0.116	0.143	0.158	0.172	0.223
RMS	0.122	0.184	0.330	0.435	0.665	0.985
MILCA 3LC	0.114	0.134	0.155	0.162	0.159	0.209
4LC	0.112	0.119	0.126	0.184	0.155	0.289
5LC	0.108	0.117	0.144	0.150	0.195	0.229
6LC	0.113	0.128	0.137	0.157	0.164	0.205
$MORAI_{\beta,\lambda}$ (f)						
MI	1.251	1.282	1.463	1.563	1.663	1.909
MICE	1.214	1.215	1.289	1.270	1.226	1.300
MISR	1.189	1.120	1.160	1.287	1.368	1.570
RMS	1.462	1.905	2.526	2.994	4.048	4.888
MILCA 3LC	1.152	1.143	1.529	1.412	1.184	1.273
4LC	1.230	1.187	1.243	1.319	1.401	1.729
5LC	1.272	1.212	1.359	1.242	1.273	1.419
6LC	1.239	1.334	1.549	1.340	1.161	1.240

(*continued on next page*)

Table 9. **M**odel **O**verall **R**elative **A**ccuracy **I**ndicator of GRM parameters (*continued*)

	5%	10%	15%	20%	25%	30%
	NMAR					
	$MORAI_{\beta}$ (g)					
MI	1.205	1.735	2.284	3.217	4.249	5.734
MICE	1.071	1.157	1.212	1.355	1.631	1.915
MISR	1.040	1.169	1.158	1.430	2.224	2.970
RMS	1.345	2.045	2.750	3.732	5.539	9.397
MILCA 3LC	0.991	0.988	1.016	1.011	1.002	0.952
4LC	1.013	1.044	1.022	1.141	1.084	0.951
5LC	1.086	1.124	1.167	1.085	1.097	0.991
6LC	1.085	1.158	1.239	1.168	1.066	0.979
	$MORAI_{\lambda}$ (h)					
MI	0.113	0.150	0.199	0.246	0.274	0.351
MICE	0.110	0.115	0.134	0.140	0.155	0.188
MISR	0.111	0.134	0.130	0.153	0.162	0.173
RMS	0.111	0.154	0.205	0.257	0.375	0.569
MILCA 3LC	0.111	0.129	0.132	0.138	0.159	0.186
4LC	0.115	0.122	0.126	0.129	0.146	0.167
5LC	0.111	0.133	0.131	0.128	0.137	0.164
6LC	0.109	0.123	0.128	0.140	0.148	0.148
	$MORAI_{\beta,\lambda}$ (i)					
MI	1.318	1.885	2.484	3.463	4.522	6.085
MICE	1.181	1.272	1.345	1.495	1.786	2.103
MISR	1.151	1.303	1.288	1.583	2.386	3.143
RMS	1.455	2.199	2.955	3.989	5.914	9.966
MILCA 3LC	1.102	1.117	1.148	1.149	1.160	1.138
4LC	1.128	1.166	1.147	1.270	1.229	1.117
5LC	1.197	1.257	1.297	1.213	1.234	1.155
6LC	1.193	1.281	1.367	1.309	1.214	1.127

Missing at Random: Under the MAR criterion the BIC mechanism would recommend selecting 5LCs as long as the percentage of missingness is medium-low (10%) and 4 LCs in the other cases (10% to 30%). According to the AIC criterion, 6 LCs are always recommended. According to the BIC criterion, a model which provides the best accuracy in terms of $MOMSE_{\lambda,\beta}$ is selected in just 3 out of 6 cases. The comparisons with the results of the other missing data imputation methods show that: i) RMS is not a valid alternative even when the rate of missingness is low (5%); ii) MISR provides the best overall accuracy in absolute and relative terms up to a certain percentage of missingness (respectively, 20% if the accuracy is measured in absolute terms using the MOMSE index and 15% if it is measured in relative terms using MORAI). MICE and MILCA show better performance with higher rates of missingness; specifically, MICE shows the best accuracy when comparisons are made in absolute terms, whereas MILCA performs better than MICE if comparisons are made in relative terms.

Under MAR conditions, more divergences in the $MOMSE_{\lambda,\beta}$ values emerge relative to the choice of the Latent Classes. However, an examination of the values of the $MORAI_{\lambda,\beta}$ index reveals these differences to be not relevant when considered in relative terms. This evidence clearly emerges if we compare the values of the $MOMSE_{\lambda,\beta}$ and $MORAI_{\lambda,\beta}$ for MILCA 6 LC and if the rate of missingness is equal to 30% (Tables 8(f) and 9(f)). Thus, in absolute terms MICE seems to perform better than MILCA (see MOMSE index) when the rate of missingness is 20% or 30%, but in two out of these three cases MILCA performs better if comparisons are made in relative terms (see the MORAI index).

A joint reading of the results of the $MOMSE_{\lambda,\beta}$ and $MORAI_{\lambda,\beta}$ shows that the BIC criterion is recommended when the dataset is affected by a low or medium level of missingness (20%) while the AIC is more suitable for the highest rates of missingness. The choice of one or the other would advocate the selection of the model with the best performances with respect to MICE in relative terms $MORAI_{\lambda,\beta}$ in 5 out 6 scenarios. However, with low rates of missingness the simulation study shows that MISR is the best choice.

Missing Not at Random: The BIC index would recommend selecting 4 LCs up to a percentage of missingness of 20% and 6 LCs for higher rates (25 and 30%). Also, with MNAR data the AIC index would recommend 6 LCs for all rates of missingness. Under the considered rates of missingness, the BIC criterion would suggest selecting the number of LCs that provides a better accuracy in the estimation of item parameters by measuring the accuracy in absolute or in relative terms. It is interesting to highlight that as the rate of missingness rises up, there is an increase in the divergence between the accuracy measures of MILCA compared to MICE (in favour of the first), on both indexes ($MOMSE_{\lambda,\beta}$, $MORAI_{\lambda,\beta}$). MI is a valid alternative to MICE and MILCA until the rate of missingness is up to 10%.

5.4.2 Simulation 2

The MILCA function was applied by selecting the number of LCs which minimizes the BIC. Tables 10 and 11 list the MOMSE and MORAI indexes under the three missing data generating processes and the six different percentages of missingness. Simulation 2 has been carried out with the main aim to assess the generalizability of the results beyond data sets used in Simulation 1. Therefore only similarities and departures from the evidences provided by Simulation 1 will be highlighted and discussed.

Missing Completely at Random: Under MCAR condition, results confirm the findings arose in Simulation 1 study. MICE and MILCA provide the highest accuracy in estimation in absolute and in relative terms up to 20% of missing values. MI and MISR are competitors of MICE and MILCA only when the rate of missingness is up to 5%.

Missing at Random: Under MAR the findings arose in Simulation 1 in terms of MOMSE are confirmed by the results gained in Simulation 2 (see Table 10). Looking at the MORAI indexes (see Table 11) it arises that MICE provides better results in terms of relative accuracy under almost all scenarios. This because comparisons in relative terms tend to highlight departures as relevant also when in absolute terms they are pointless: e.g. if a parameter is estimated 0.03 instead of 0.01, in relative terms its weight in the MORAI function is 3.

Missing Not at Random: Under MNAR all the evidences arose in Simulation 1 are confirmed. MILCA is selected as the best imputation method in terms of absolute and relative accuracy for rates of missingness from 10% up to 30% (see Table 11). Under the MNAR condition, the following three pieces of evidences emerge from the simulation study (i) MILCA seems to perform better than any other imputation procedure considered, whatever is the criterion to measure the accuracy (absolute or relative) might be; (ii) the BIC criterion ensures the selection of the number of LCs which provide the best accuracy in estimation for both indexes (MOMSE and MORAI) and for any rate of missingness; (iii) the RMS and MI method are not a valid alternative even with low-medium rates of missingness (10-15%).

6. Discussion

This article assesses the performances of an *ad hoc* multiple imputation approach for categorical items based on LCA and measures its accuracy in the IRT framework with respect to other imputation methods. The proposed procedure uses functions implemented for LCA to detect unobserved categorical unordered latent classes characterised by different vectors of item response probabilities and to assign each individual to one of them. The vectors of parameters of each class is then used to sample plausible values for imputation purposes. As a result multiple datasets are created which are then analyzed with standard MIA tools.

The accuracy of the procedure was validated for the estimation of the item parameters of Graded Response Models by simulating missing data according to different mechanisms in two benchmark datasets. For this aim two simulation studies have been carried out. Both simulation designs are also devised to validate the effectiveness of MILCA with regard to other single and multiple imputation methods, under ignorable and non ignorable missing data generating processes and under increasing percentages of missingness. The procedure was validated by advancing a set of measures which allow us to summarise the accuracy of the location and discrimination parameters in an overall index suitable to highlighting the procedure effectiveness in absolute and relative terms.

Table 10. **M**odel **O**verall **M**ean **S**quare **E**rror of GRM parameters

	5%	10%	15%	20%	25%	30%
MCAR						
$MOMSE_{\beta}$ (a)						
MI	0.853	0.962	1.118	1.300	1.361	1.672
MICE	0.839	0.873	0.888	0.906	0.875	0.929
MISR	0.841	0.851	0.843	0.860	1.085	1.269
RMS	1.101	1.590	2.017	2.697	3.591	4.964
MILCA 7 LC	0.870	0.895	0.901	0.916	0.959	0.979
$MOMSE_{\lambda}$ (b)						
MI	0.057	0.080	0.102	0.142	0.206	0.292
MICE	0.049	0.061	0.086	0.091	0.099	0.105
MISR	0.050	0.065	0.091	0.161	0.323	0.460
RMS	0.127	0.381	0.672	1.232	2.054	3.095
MILCA 7 LC	0.052	0.064	0.071	0.080	0.115	0.123
$MOMSE_{\beta,\lambda}$ (c)						
MI	0.910	1.042	1.220	1.442	1.567	1.964
MICE	**0.888**	0.934	0.974	0.998	**0.974**	**1.034**
MISR	0.891	0.916	0.934	1.021	1.408	1.729
RMS	1.228	1.971	2.688	3.929	5.645	8.059
MILCA 7 LC	0.922	**0.958**	**0.972**	**0.996**	1.074	1.102
MAR						
$MOMSE_{\beta}$ (d)						
MI	0.867	0.940	1.028	1.264	1.843	2.061
MICE	0.868	0.893	0.914	0.888	0.971	0.935
MISR	0.883	0.894	1.001	1.054	1.153	1.412
RMS	1.330	1.883	2.449	3.351	5.151	6.740
MILCA 7 LC	0.866	0.910	0.937	0.920	1.195	1.309
$MOMSE_{\lambda}$ (e)						
MI	0.073	0.086	0.132	0.175	0.207	0.309
MICE	0.056	0.063	0.077	0.086	0.091	0.095
MISR	0.074	0.092	0.144	0.202	0.282	0.431
RMS	0.216	0.532	0.890	1.590	2.631	3.798
MILCA 7 LC	0.053	0.067	0.067	0.075	0.168	0.212
$MOMSE_{\beta,\lambda}$ (f)						
MI	0.939	1.026	1.160	1.439	2.050	2.370
MICE	0.924	**0.956**	**0.992**	**0.974**	**1.062**	**1.030**
MISR	0.958	0.986	1.144	1.256	1.435	1.844
RMS	1.547	2.415	3.339	4.940	7.782	10.537
MILCA 7 LC	**0.919**	0.977	1.005	0.995	1.363	1.521
MNAR (g)						
$MOMSE_{\beta}$						
MI	0.886	0.986	1.159	1.278	1.565	1.912
MICE	0.871	0.927	0.990	1.071	1.160	1.406
MISR	0.837	0.851	0.993	1.259	1.548	2.026
RMS	1.349	1.791	2.657	3.688	4.867	7.251
MILCA 7 LC	0.852	0.886	0.892	0.914	0.882	1.010
$MOMSE_{\lambda}$ (h)						
MI	0.053	0.060	0.109	0.093	0.132	0.187
MICE	0.048	0.065	0.075	0.096	0.137	0.131
MISR	0.050	0.062	0.120	0.251	0.361	0.710
RMS	0.202	0.347	0.707	1.198	1.801	2.710
MILCA 7 LC	0.052	0.067	0.095	0.086	0.076	0.117
$MOMSE_{\beta,\lambda}$ (i)						
MI	0.939	1.047	1.269	1.372	1.698	2.099
MICE	0.919	0.992	1.066	1.167	1.297	1.537
MISR	0.887	0.913	1.113	1.510	1.910	2.736
RMS	1.551	2.138	3.364	4.886	6.668	9.961
MILCA 7 LC	**0.904**	**0.953**	**0.988**	**0.999**	**0.958**	**1.128**

Number of Latent Classes selected with the BIC criterion

Table 11. **M**odel **O**verall **R**elative **A**ccuracy **I**ndicator of GRM parameters

	5%	10%	15%	20%	25%	30%
			MCAR			
			$MORAI_{\beta}$ (a)			
MI	1.835	6.459	16.443	32.751	46.586	58.481
MICE	0.914	0.995	1.158	1.046	1.164	1.611
MISR	0.900	1.216	2.494	3.322	5.974	9.527
RMS	4.604	14.738	31.372	45.185	63.215	67.797
MILCA 7 LC	1.123	1.003	1.040	1.051	1.205	1.440
			$MORAI_{\lambda}$ (b)			
MI	0.054	0.069	0.093	0.120	0.142	0.197
MICE	0.049	0.058	0.079	0.085	0.090	0.091
MISR	0.050	0.056	0.071	0.112	0.209	0.296
RMS	0.143	0.434	0.796	1.482	2.444	3.610
MILCA 7 LC	0.051	0.056	0.066	0.069	0.101	0.107
			$M0RAI_{\beta,\lambda}$ (c)			
MI	1.889	6.528	16.536	32.871	46.728	58.678
MICE	**0.963**	**1.052**	1.237	1.132	**1.254**	1.702
MISR	0.949	1.272	2.565	3.434	6.183	9.823
RMS	4.746	15.172	32.167	46.667	65.660	71.407
MILCA 7 LC	1.174	1.059	**1.105**	**1.119**	1.305	**1.547**
			MAR			
			$MORAI_{\beta}$ (d)			
MI	3.120	7.660	20.805	41.683	67.534	88.869
MICE	1.043	1.697	2.510	1.309	1.249	1.473
MISR	1.154	3.302	4.987	7.229	11.535	22.372
RMS	6.303	14.755	25.686	37.168	48.754	67.337
MILCA 7 LC	1.076	3.350	3.055	2.721	3.271	1.466
			$MORAI_{\lambda}$ (e)			
MI	0.056	0.069	0.086	0.129	0.177	0.219
MICE	0.052	0.059	0.072	0.077	0.080	0.091
MISR	0.060	0.072	0.112	0.132	0.182	0.279
RMS	0.223	0.599	1.034	1.871	2.968	4.279
MILCA 7 LC	0.051	0.062	0.061	0.068	0.147	0.183
			$MORAI_{\beta,\lambda}$ (f)			
MI	3.177	7.729	20.891	41.812	67.711	89.088
MICE	1.095	1.756	2.581	1.386	1.330	1.564
MISR	1.214	3.375	5.099	7.362	11.717	22.651
RMS	6.526	15.353	26.720	39.039	51.723	71.617
MILCA 7 LC	1.128	3.412	3.117	2.789	3.418	1.650
			NMAR			
			$MORAI_{\beta}$ (g)			
MI	3.366	4.647	6.994	15.311	35.922	36.782
MICE	0.979	1.889	4.228	7.069	8.211	13.771
MISR	1.110	2.424	3.231	6.364	17.821	19.653
RMS	7.530	11.989	26.675	50.423	78.909	77.339
MILCA 7 LC	1.214	1.568	1.673	1.319	2.879	1.993
			$MORAI_{\lambda}$ (h)			
MI	0.051	0.057	0.088	0.075	0.104	0.143
MICE	0.048	0.061	0.068	0.087	0.109	0.126
MISR	0.048	0.054	0.090	0.181	0.262	0.519
RMS	0.217	0.404	0.836	1.422	2.180	3.208
MILCA 7LC	0.052	0.069	0.073	0.081	0.074	0.097
			$MORAI_{\beta,\lambda}$ (i)			
MI	3.417	4.704	7.082	15.386	36.026	36.925
MICE	1.028	1.951	4.295	7.156	8.320	13.897
MISR	1.157	2.478	3.321	6.545	18.083	20.172
RMS	7.747	12.392	27.511	51.845	81.089	80.547
MILCA 7LC	1.266	1.637	1.747	1.400	2.953	2.090

The two simulation studies show that the MILCA procedure is a valid imputation method for carrying out analysis in Item Response Theory framework whenever missing data arise according to different generating processes. All the results agree in demonstrating that MICE and MILCA seem to be competing imputation procedures if the aim is to maximize the accuracy in absolute and relative terms and if the data set is affected by a high rate of missingness, whatever the missing data generating process might be. None of the two imputation methods emerge as superior under all simulated conditions under ignorable missing data processes. As regards the aim of validating under which conditions MILCA's performance is superior in comparison with the other imputation methods, a special focus is devoted in Simulation 1 to assess divergences in the accuracy of the results, by selecting the optimal number of latent classes according to different goodness of fit criteria. Furthermore, results show that MILCA performs quite well when the missing data mechanism is not ignorable whatever the method to measure the accuracy might be.

To sum up, the simulation study shows that MICE and MILCA seem to be interchangeable procedures under MCAR conditions. Under MCAR the BIC criterion proved to be the best index for selecting the optimal number of latent classes using MILCA. Under MAR conditions, the main difference which arises regarding the bias in relative terms (using the MORAI index) is that in Simulation 1 MILCA seems to provide more accurate estimates of the threshold and discrimination parameters when the rate of missingness is severe (20% or more), whereas in Simulation 2, MICE seems to be superior. Specifically, also under MAR none of the two methods emerges as superior in both simulation studies (Simulation 1 and Simulation 2) under all the rates of missing values. The real advantage of using MILCA is detectable with the MNAR scenario, given that in the latter scenario, the accuracy of MILCA is higher than MICE when MCAR applies. This evidence clearly emerges by comparing the values of the MOMSE and MORAI indexes for item-category location and discrimination parameters listed in Tables 8(g,h,i), 9(g,h,i), 10(g,h,i), 11(g,h,i).

The proposed method has the advantage of being easy to use with any categorical set of item measured on dichotomous, nominal or Likert-Type scales by using the function `mipoLCA` (available in the supplementary materials online) which recalls scripts already implemented for LCA in `R` and uses them for imputation purposes. Its good performances with respect to the other missing data handling methods under non ignorable missingness conditions make the procedure the most convenient choice to be adopted even when the missing data mechanism is not detectable.

Further research aim to extend the MILCA procedure in order to use the information provided by individuals’ covariates to predict the latent class membership probabilities and to assess the robustness of the procedure under different missing data mechanisms. The implementation and validation of a multiple imputation approach based on Latent Class Regression Analysis would allow us to maximize the use of the information available in the dataset in order to predict non responses.

References

AGRESTI, A. (2002), *Categorical Data Analysis*, Hoboken: Wiley-Interscience.

AKE, C.F. (2005), “Rounding After Multiple Imputation with Non-Binary Categorical Covariates”, paper presented at the annual meeting of the *SAS User Group International, Philadelphia*.

BAKER, F.B., and KIM, S.H. (2004), *Item Response Theory: Parameter Estimation Techniques*, New York: Dekker.

BARALDI, A.N., and ENDERS, C.K. (2010), “An Introduction to Modern Missing Data Analyses”, *Journal of School Psychology, 48,* 5–37.

BERNAARDS, C.A., and SIJTSMA, K. (1999), “Factor Analysis of Multidimensional Polytomous Item Response Data from Ignorable Item Non Response”, *Multivariate Behavioral Research, 34,* 277–314.

BIRNBAUM, A. (1968), “Statistical Theories of Mental Test Scores”, in *Some Latent Trait Models and Their Use in Inferring an Examinee’S Ability*, eds F.M. Lord and M.R. Novick, Reading: Addsion-Wesley, pp. 395–497.

CARPITA, M., and MANISERA, M. (2011), “On the Imputation of Missing Data in Surveys with Likert-Type Scales”, *Journal of Classification, 28,* 93–112,

EDELEN, M.O., and REEVE, B.B. (2007), “Applying Item Response Theory (IRT) Modeling to Questionnaire Development, Evaluation and Refinement”, *Quality of Life Researches, 16*, 5–18

ENDERS, G.K. (2004), “The Impact of Missing Data on Sample Reliability Estimates: Implications of Reliability Reporting Practices”, *Educational and Psychological Measurement, 64(3)*, 419–436,

FINCH, H. (2008), “Estimation of Item Response Theory Parameters in the Presence of Missing Data”, *Journal of Educational Measurement, 45(3)*, 225–245.

FINCH, H. (2010), “Imputation Methods for Missing Categorical Questionnaire Data: A Comparison of Approaches”, *Journal of Data Science, 8(8),* 361–378.

FINCH, H. (2011), “The Impact of Missing Data on the Detection of Nonuniform Differential Item Functioning”, *Educational and Psychological Measurement, 71(4),* 663–683.

HUISMAN, M. (1999), *Item Nonresponse: Occurence, Causes, and Imputation of Missing Answers to Test Items*, Leiden, The Netherlands: DSWO Press.

LINZER, D.A., and LEWIS, J. (2011), “poLCA: Polytomous Variable Latent Class Analysis”, *Journal of Statistical Software, 42(10)*, 1–29.

LITTLE, R.J.A., and RUBIN, D.B. (2002), *Statistical Analysis with Missing Data* (2nd. ed.), New York: John Wiley.

MULLIS, I.V.S., MARTIN, M.O., FOY, P., and DRUCKER, K.T. (eds.) (2012), “PIRLS 2011 International Results in Reading”, *2012 International Association for the Evaluation of Educational,* Chestnut Hill MA: TIMSS & PIRLS International Study Center Boston College.

NYLUND, K., ASPAROUHOV, T., and BENGT, O.M. 2007, “Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study”, *Structural Equation Modeling, 14(4)*, 535–569.

RAAIJMAKERS, A.W. (1999), “Effectiveness of Different Missing Data Treatments in Surveys with Likert-Type Data: Introducing the Relative Mean Substitution Approach”, *Educational and Psychological Measurement, 59(5),* 725–748.

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”, *Survey Methodology, 27,* 85–95.

RIZOPOULOS, D. (2006), “ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses”, *Journal of Statistical Software, 17(5)*, 1–25.

RUBIN, D. (1976), “Inference and Missing Data”, *Biometrika, 63,* 581–592.

SAMEJIMA, F. (1969), “Estimation Of Ability Using a Response Pattern of Graded Scores”, *Psychometrika Monograph, 17*.

SCHAFER, J. (1997), *Analysis of Incomplete Multivariate Data*, London: Chapman and Hall.

SCHAFER, J., and GRAHAM, J.W. (2002), “Missing Data: Our View of the State of the Art”, *Psychological Methods, 7(2)*, 147–177.

SIJTSMA, K., and VAN DER ARK, L.A., (2003), “Investigation and Treatment of Missing Item Scores in Test and Questionnaire Data”, *Multivariate Behavioral Research, 38(4)*, 505–528.

SULIS, I. (2013), “A Further Proposal to Perform Multiple Imputation on a Bunch of Polythomous Items based on Latent Class Analysis”, in *Statistical Models for Data Analysis: Studies in Classification, Data Analysis, and Knowledge Organization*, eds. P. Giudici, S. Ingrassia, and M. Vichi, Heidelberg: Springer-Verlag.

SULIS, I., and PORCU, P. (2008), “Assessing the Effectiveness of a Stochastic Regression Imputation Method for Ordered Categorical Data”, *CRENoS Working Papers, 4.*

VAN BUUREN, S., and OUDSHOORN, C.G.M. (2011), “MICE: Multivariate Imputation by Chained Equations”, *Journal of Statistical Software, 45(3),* 1–67.

VERMUNT, J.K, VAN GINKEL, J.R., VAN DER ARK, L.A., and SIJTSMA, K. (2008), “Multiple Imputation of Categorical Data Using Latent Class Analysis”, *Sociological Methodology, 33,* 269–297.

WU, W., JIA, F., and ENDERS, C. (2015), “A Comparison of Imputation Strategies for Ordinal Missing Data on Likert Scale Variables”, *Multivariate Behavioral Research, 50,* 484–503.