Journal of Data Science and Modeling, Vol. 3, No. 2, 27-43, June 2025

Research Manuscript

Bayesian Analysis of Missing Data and Its Application in Categorical Data of TIMSS Test

Zahra Ghaffari Irdmousa*[1]

[1]Department of Statistics, Allameh Tabataba'i University, Tehran, Iran

Recieved: 06/04/2026 Accepted: 29/07/2026

Abstract:

In this article, unlike recently developed methods that can only distinguish associations between pairs of variables, the Bayesian Multilevel Latent Class (BMLC) model for the multiple imputation of nested categorical data is considered. This model is flexible enough to automatically deal with complex interactions in the joint distribution of the variables to be estimated. After presenting the model, I run it on a real dataset, namely the mathematics test of TIMSS 2015, which inherently contains missing data. Ultimately, the completed data for both levels (level 1: students, and level 2: schools) are obtained. To assess the performance of the BMLC model, we compare it with listwise deletion (LD) using R software. Conclusion: The results show that the BMLC model is reliable and capable of covering the Bayesian estimator with lower risk in this research.

Keywords: test of TIMSS, Bayesian analysis, missing data, categorical data, Bayesian Multilevel Latent Class model.

Mathematics Subject Classification (2010): 62H25, 62H35.

*Corresponding Author: Z.ghaffari85@yahoo.com

1. Introduction

Statistical inference in the presence of missing data is one of the most critical challenges in empirical research. Missing data can reduce the statistical power of a study, decrease the efficiency of estimators, and introduce systematic bias (Rashidi-Nejad and Navabpour, 2010). Rubin (1976) conceptualized the mechanisms of missing data as follows: when missingness is independent of both observed and unobserved data, it is classified as Missing Completely at Random (MCAR). If the missingness of a variable depends only on observed values (for instance, missingness in a second-round response depending on the first-round observed response), it is classified as Missing at Random (MAR). Finally, if the probability of missingness depends on the unobserved values themselves, it is referred to as Missing Not at Random (MNAR) (Rubin, 1976).

In general, missingness manifests in two primary forms:

1– Unit nonresponse, where all variables for a given sample unit are missing,
2– Item nonresponse, where only specific variables/items for a sample unit are missing.

One of the most effective approaches to address missingness is the imputation of missing values. Since missing data are practically unavoidable in empirical applications, developing and evaluating robust imputation methods serves as the primary motivation for this study.

Data from the TIMSS[1] assessment exhibit an MAR mechanism. The missing observations in this study are classified as item nonresponse. In previous studies conducted on TIMSS datasets, multiple imputation has not yet been systematically applied to address missingness, a gap that the current research aims to bridge. Specifically, this paper utilizes the Bayesian Multilevel Latent Class (BMLC)[2] model, designed for the imputation of nested missing data.

Nested or multilevel data structures are common in the educational, social, and medical sciences. In such frameworks, Level 1 (lower-level) units, such as students, citizens, or patients, are nested within Level 2 (higher-level) units, such as schools, cities, or hospitals. When lower-level units within the same cluster exhibit correlation, the nested structure of the data must be explicitly accounted for. Although standard single-level analysis assumes independent Level 1 observations, multilevel modeling incorporates these dependencies to prevent biased estimates and incorrect standard errors (Van Buuren and et al., 2011; Vidotto et al., 2018).

[1] Trends in International Mathematics & Science Study (TIMSS)
[2] Bayesian Multilevel Latent Class Model (BMLC)

2. Literature Review

Research on missing data imputation is broadly categorized into two paradigms:

1– Bayesian imputation,
2– Classical (frequentist) imputation.

2.1 Bayesian Imputation

The application of Bayesian methods to missing data that are missing at random (MAR) in multivariate analysis was pioneered by Rubin (1976). Subsequently, Rubin (1987) proposed multiple imputation (MI) to address missing data issues. In this approach, rather than replacing a missing value with a single estimate, multiple values are generated to fill the gaps in the dataset, thereby capturing the uncertainty associated with the missing values.

Di Zio et al. (2004) demonstrated the utility of Bayesian networks for missing value imputation. They argued that the primary advantages of utilizing Bayesian networks as imputation models lie in their capacity to preserve complex statistical relationships among variables and their suitability for handling missingness in high-dimensional datasets.

An alternative framework for estimating both the joint probability distribution and the marginal probability of variables containing missing data—by treating the missing value indicator as an unknown parameter and using Markov Chain Monte Carlo (MCMC) methods—was developed by Gilks and Roberts (1996).

In a study titled "Bayesian Multilevel Latent Class Models for the Multiple Imputation of Nested Categorical Data," Vidotto et al. (2018) demonstrated that multilevel mixture models can recover unbiased parameter estimates for analysis models. Furthermore, this approach accurately reflects the uncertainty stemming from missing data and outperforms competing imputation techniques.

Similarly, Vidotto et al. (2020) proposed the Bayesian Mixture Latent Markov (BMLM) model for the MI of longitudinal categorical data. Through simulation studies and empirical testing, the authors evaluated the BMLM method against complete-case analysis and Multivariate Imputation by Chained Equations (MICE)[3]. Their findings indicated that the BMLM method performs robustly in recovering the parameters of interest in the analysis model.

Additionally, Ahmadi-Kahjouq (2014) evaluated missing data imputation using Bayesian classification. Their simulation results showed that Bayesian network models compete effectively with classical methods and exhibit superior performance on specific datasets.

[3]Multivariate Imputation by Chained Equations (MICE)

In another comparative study, Karimlou and et al. (2006) examined Bayesian and classical methods for estimating the parameters of logistic regression models with missing values in auxiliary variables. They noted that when the auxiliary variables X are missing at random, the likelihood function for the observed data is well-defined and can be analyzed as if the data were complete. Using this likelihood framework, they compared classical maximum likelihood estimation with Bayesian estimation via MCMC. The estimates obtained from the Bayesian MCMC approach were closer to the complete-data standard estimates than those derived from classical estimation methods.

For a comprehensive review of classical imputation methodologies, the reader is referred to Ghaffari Irdmousa (2020).

As a highly flexible framework in the literature, MI (Rubin, 1987) uses an imputation model solely to draw imputed values from the posterior predictive distribution of the missing data given the observed data. Following this step, standard complete-data analysis can be performed on each of the M completed datasets. This procedure ensures that uncertainty arising from the sampling stage is distinguished from the uncertainty introduced by the imputation step in the pooled estimates and their standard errors. A key advantage of MI is that, once the imputation step is finalized, any standard statistical analysis can be applied to the completed datasets (Allison, 2009; Vidotto et al., 2018). Consequently, the advantages of MI can be leveraged whether the data are missing completely at random (MCAR), missing at random (MAR), or even missing not at random (MNAR) (Vidotto et al., 2018).

Two major paradigms for specifying imputation models are Full Conditional Specification (FCS) and Joint Modeling (JM, often implemented via packages like JOMO). The former is fundamentally a variable-by-variable approach that requires specifying separate conditional models for each variable with missing values. In contrast, the latter requires specifying a single joint multivariate distribution for all variables from which the imputations are drawn.

Most standard software packages, such as the `mice` package in R (which primarily relies on FCS), do not readily accommodate the MI of multilevel categorical data. FCS approaches for multilevel data typically struggle when categorical variables have more than two categories and have not yet been fully extended to the imputation of Level 2 predictors. Conversely, the JOMO framework operates under a Bayesian paradigm and utilizes Gibbs sampling for imputations. Despite its advantages, JOMO has limitations: it assumes multivariate normality, meaning that higher-order associations, such as interactions and nonlinear relationships, are often neglected. This makes JOMO less flexible and potentially leads to suboptimal imputations when complex dependencies are present in the data.

Given the critical impact of missing data on statistical inference, and noting that multiple imputation has not yet been applied to the TIMSS assessment data, this study implements a multiple imputation framework based on Bayesian analysis. Specifically, this research aims to achieve the following objectives:

1– To mitigate the bias introduced by missing observations in the nested categorical dataset of the TIMSS test using multiple imputation.
2– To compare the parameter estimates, standard errors, and P-values obtained from listwise deletion (LD) against those obtained from the BMLC model.

3. Methodology

One of the primary objectives of this study is to apply Bayesian methods to analyze TIMSS assessment data while accounting for the underlying missingness structure. Specifically, this research utilizes the Bayesian Multilevel Latent Class (BMLC) model—a multilevel mixture framework—for missing data imputation, using empirical data from the 2015 TIMSS administration in Iran. The adoption of this methodology is motivated by its high flexibility and its suitability for the nested structure of TIMSS datasets. Because this study utilizes secondary data from an established international assessment, it is classified as a secondary data analysis.

The target population and sample of interest consist of eighth-grade students (referred to as Population 2 in TIMSS) who participated in the TIMSS 2015 mathematics assessment in Iran. The sampling design of TIMSS is a two-stage stratified cluster sampling design, where schools (sampling units) are selected with probability proportional to size (PPS) at the first stage, and classrooms are sampled at the second stage (Ghaffari Irdmousa, 2014).

The instrument analyzed is the TIMSS 2015 eighth-grade mathematics test, which comprises 14 booklets distributed across student samples. The total sample size selected in Iran was reported as 6,130 students. For the purposes of this study, Booklet 1 (corresponding to Block 1) was selected for analysis. This block contains data for $N = 440$ students across 35 mathematics items. Multiple imputation is performed on this dataset using R software.

Much of the theoretical framework in the BMLC section and the details in Algorithm 1 are adapted from or build upon the methodology proposed by Vidotto et al. (2018).

3.1 The BMLC Model for Multiple Imputation (MI)

The BMLC imputation model proposed in this study corresponds to the nonparametric version of the multilevel LC model introduced by Vermunt (2003) in a

frequentist framework. The BMLC model is capable of capturing heterogeneity at both hierarchical levels of the dataset. This is achieved by clustering Level 2 units into Level 2 latent classes; conditional on these clusters, Level 1 units are further classified into Level 1 latent classes.

Let $\mathbf{D} = \mathbf{D}^{\text{obs}} = (\mathbf{Z}, \mathbf{Y})$ denote a nested dataset with J Level 2 units and n_j Level 1 units nested within each Level 2 unit j ($j = 1, \ldots, J$). Assume that the dataset contains T Level 2 categorical variables $Z_1, \ldots, Z_t, \ldots, Z_T$, where each variable Z_t has R_t observed categories ($t = 1, \ldots, T$). Similarly, let there be S Level 1 categorical variables $Y_1, \ldots, Y_s$, where each variable Y_s has U_s observed categories ($s = 1, \ldots, S$).

We denote the vector of the T Level 2 item scores for Level 2 unit j by $\mathbf{z_j} = (z_{j1}, \ldots, z_{jT})$, and the collective vector of Level 1 observations within Level 2 unit j by $\mathbf{y_j} = (\mathbf{y_{j1}}, \ldots, \mathbf{y_{ji}}, \ldots, \mathbf{y_{jn_j}})$, where $\mathbf{y_{ji}} = (y_{ji1}, \ldots, y_{jiS})$ represents the vector of the S Level 1 item scores for Level 1 unit i within Level 2 unit j. Let W_j represent the Level 2 latent class variable, which takes L classes ($l = 1, \ldots, L$), and let $X_{ji} \mid W_j = l$ represent the Level 1 latent class variable, which takes K classes ($k = 1, \ldots, K$) within the l-th Level 2 latent class.

The joint probability model for the higher-level observations can be expressed as:

$$\text{pr}(\mathbf{Z}_j = \mathbf{z}_j, \mathbf{Y}_j = \mathbf{y}_j) = \sum_{l=1}^{L} \text{pr}(W_j = l) \prod_{t=1}^{T} \text{pr}(Z_{jt} = z_{jt} \mid W_j = l) \prod_{i=1}^{n_j} \text{pr}(\mathbf{Y}_{ji} = \mathbf{y}_{ji} \mid W_j = l)$$

where the conditional probability model for the lower-level observations is defined as:

$$\text{pr}(\mathbf{Y}_{ji} = \mathbf{y}_{ji} \mid W_j = l) = \sum_{k=1}^{K} \text{pr}(X_{ji} = k \mid W_j = l) \prod_{s=1}^{S} \text{pr}(Y_{jis} = y_{jis} \mid W_j = l, X_{ji} = k)$$

Figure 1 illustrates the fundamental graphical model of this hierarchical structure. As shown in Figure 1, the model allows the number of Level 1 latent variables to vary across Level 2 units (indexed by j), while simultaneously showing how the Level 2 latent class variable W_j influences Z_j, X_{ji}, and Y_{ji}.

As in standard latent class analysis, we assume multinomial distributions for the Level 1 latent class variable $X \mid W$ and the conditional response distributions $\text{pr}(Y_s \mid W, X)$. Furthermore, we assume multinomial distributions for the conditional responses at the higher level $\text{pr}(Z_t \mid W)$. Since we employ the nonparametric version of the multilevel LC model, the Level 2 mixture variable W is also assumed to follow a multinomial distribution.

The Level 2 mixture variable W is supposed to follow a multinomial distribu-

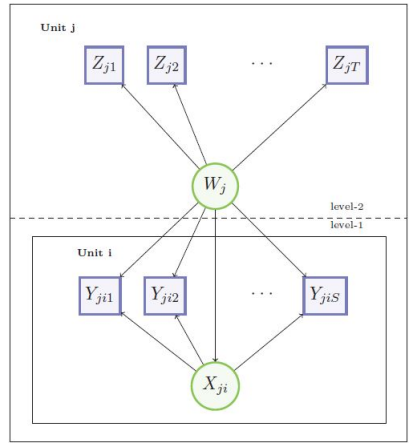

Figure 1: Graphical demonstration of the multilevel latent class model

tion.

$$W \sim \text{Multinom}(\pi_W),$$
$$X \mid W = l \sim \text{Multinom}(\pi_{lX}), \quad l = 1, \ldots, L,$$
$$Z_t \mid W = l \sim \text{Multinom}(\pi_{lt}), \quad t = 1, \ldots, T, \quad l = 1, \ldots, L,$$
$$Y_s \mid W = l, X = k \sim \text{Multinom}(\pi_{lks}), \quad s = 1, \ldots, S, \quad l = 1, \ldots, L, \quad k = 1, \ldots, K$$

A vector including the probabilities of each category of the corresponding multinomial distribution is indicated by the parameters.

$$\pi_W = (\pi_1, \ldots, \pi_l, \ldots, \pi_L)$$
$$\pi_{lX} = (\pi_{l1}, \ldots, \pi_{lk}, \ldots, \pi_{lK})$$
$$\pi_{lt} = (\pi_{lt1}, \ldots, \pi_{ltr}, \ldots, \pi_{ltR_t})$$
$$\pi_{lks} = (\pi_{lks1}, \ldots, \pi_{lksu}, \ldots, \pi_{lksU_s})$$

The whole parameter vector is: $\pi = (\pi_w, \pi_{lx}, \pi_{lt}, \pi_{lks})$. Supposing multinomial distributions for every item of the model, we can rewrite the model:

$$\text{pr}(\mathbf{Z}_j = \mathbf{z}_j, \mathbf{Y}_j = \mathbf{y}_j; \pi) = \sum_{l=1}^{L} \pi_l \prod_{t=1}^{T} \prod_{r=1}^{R_t} (\pi_{ltr})^{I_{jt}^r} \prod_{i=1}^{n_j} \pi_{jil} \tag{3.1}$$

in which $I_{jt}^r = 1$ if $z_{jt} = r$, and 0 otherwise, and $\pi_{jil} = \text{pr}(\mathbf{Y_{ji}} = \mathbf{y_{ji}} \mid W_j = l)$. The recent quantity is derived from the lower-level data model, stated by:

$$\pi_{jil} = \sum_{k=1}^{K} \pi_{lk} \prod_{s=1}^{S} \prod_{u=1}^{U_s} (\pi_{lksu})^{I_{jis}^u} \tag{3.2}$$

where $I^u_{jis} = 1$ if $y_{jis} = u$ and 0 otherwise.

To gain a Bayesian estimation of the model described by Equations (3.1) and (3.2), a prior distribution for p is required. For the multinomial distribution, a class of conjugate priors is the Dirichlet distribution. For the BMLC model, we suppose as priors:

$$\begin{aligned} (a) \qquad & \pi_w \sim \mathrm{Dir}(\alpha_w) \\ (b) \qquad & \pi_{lX} \sim \mathrm{Dir}(\alpha_{lX}) \\ (c) \qquad & \pi_{lt} \sim \mathrm{Dir}(\alpha_{lt}) \\ (d) \qquad & \pi_{lks} \sim \mathrm{Dir}(\alpha_{lks}) \end{aligned}$$

Vectors are indicated by the hyperparameters of the Dirichlet distribution. α_w corresponds to the vector $(\alpha_1, \ldots, \alpha_l, \ldots, \alpha_L)$

$$\begin{aligned} \alpha_{lX} &= (\alpha_{l1}, \ldots, \alpha_{lk}, \ldots, \alpha_{lK}) \qquad \forall l \\ \alpha_{lt} &= (\alpha_{lt1}, \ldots, \alpha_{ltr}, \ldots, \alpha_{ltR_t}) \qquad \forall l, t \\ \alpha_{lks} &= (\alpha_{lks1}, \ldots, \alpha_{lksu}, \ldots, \alpha_{lksu_s}) \qquad \forall l, k, s \end{aligned}$$

The vector including every hyperparameter value will be denoted by

$$\alpha = (\alpha_w, \ldots, \alpha_{lks}) \qquad \forall l, k, s, t.$$

Reducing the hyperparameter of the items' conditional distribution probabilities to .01 (or .05) directed the imputation model to gain unbiased terms.

Model selection can be done similarly to Gelman et al. (2013) method. The process needs implementing the Gibbs sampler described in Algorithm 1 without pace 7 with large L^* and K^* and putting hyperparameters for the LC probabilities that can support empty extra components. These values could be equal to $\alpha_l = 1/L^*$ and $\alpha_k = 1/K^*$. At the end of every iteration of the elementary Gibbs sampler, we keep track of the number of LCs that are assigned in order to gain a distribution for L and K when the algorithm ends. If the posterior maxima $L_{\max}$ and $K_{\max}$ of such distributions are smaller than the suggested L^* and K^*, in the next pace, the imputations can be carried out with $L_{\max}$ and $K_{\max}$. Nevertheless, if either $L_{\max}$ or $K_{\max}$ (or both of them) equals L^* or K^*, we rerun the elementary Gibbs sampler by raising the corresponding value(s) and repeat the process until optimal L and K are found.

This imputation method is examined for BMLC models in real data of the TIMSS test in this research.

In the first section of Algorithm 1, the BMLC model is estimated first by allocating the units to the LCs (paces 1 and 2) via the posterior membership probabilities and next by updating the model parameter (paces 3–6). At the end

of the Gibbs sampler (pace 7), after the model has been estimated, we impute the missing data via M draws from $\text{pr}(\pi \mid \mathbf{Y}^{\text{obs}}, \mathbf{Z}^{\text{obs}})$.

After fixing K, L, and α, we must determine I, the total number of iterations for the Gibbs sampler. If we indicate the number of the iterations applied for the burn-in with b, we will allocate I, such that $I = b+(I-b)$, where $I-b$ is the number of iterations necessary for the estimation of the distribution $\text{pr}(\pi \mid \mathbf{Y}^{\text{obs}}, \mathbf{Z}^{\text{obs}})$, from which we will draw the parameter values necessary for the imputations.

Algorithm 1.

(A) Part 1. For $h = 1, \ldots, I$:

1– For $j = 1, \ldots, J$ sample $W_j^{(h)} \in \{1, \ldots, L\}$ from a Multinomial distribution with the posterior membership probabilities at level two as parameters (and sample size 1), calculated via:

$$\text{pr}\left(W_j^{(h)} = l \mid \mathbf{Y}_j^{\text{obs}}, \mathbf{Z}_j^{\text{obs}}, \pi^{(h-1)}\right) = \frac{\pi_l^{(h-1)} \left\{ \prod_{t=1}^{T} \prod_{r=1}^{R_t} \left(\pi_{ltr}^{(h-1)}\right)^{I_{jl}^{r*}} \right\} \left\{ \prod_{i=1}^{n_j} \sum_{k=1}^{K} \pi_{lk}^{(h-1)} \prod_{s=1}^{S} \prod_{u=1}^{U_s} \left(\pi_{lksu}^{(h-1)}\right)^{I_{jis}^{u*}} \right\}}{\sum_{p=1}^{L} \pi_p^{(h-1)} \left\{ \prod_{t=1}^{T} \prod_{r=1}^{R_t} \left(\pi_{ptr}^{(h-1)}\right)^{I_{jt}^{r*}} \right\} \left\{ \prod_{i=1}^{n_j} \sum_{k=1}^{K} \pi_{pk}^{(h-1)} \prod_{s=1}^{S} \prod_{u=1}^{U_s} \left(\pi_{pksu}^{(h-1)}\right)^{I_{jis}^{u*}} \right\}},$$

in which $I_{jt}^{r*} = 1$ if $Z_{jt} = r$ and $Z_{jt} \in \mathbf{Z}^{\text{obs}}$ or $I_{jt}^{r*} = 0$ otherwise, and similarly $I_{jis}^{u*} = 1$ if $Y_{jis} = u$ and $Y_{jis} \in \mathbf{Y}^{\text{obs}}$ or $I_{jis}^{u*} = 0$ otherwise;

2– For $i = 1, \ldots, n_j$, for all j, and given $W_j^{(h)}$, sample $X_{ji}^{(h)} \in \{1, \ldots, K\}$ from a Multinomial distribution with the posterior membership probabilities at level one as parameter (and sample size 1), calculated via:

$$\text{pr}\left(X_{ji}^{(h)} = k \mid W_j^{(h)} = l, \mathbf{Y}_j^{\text{obs}}, \mathbf{Z}_j^{\text{obs}}, \pi^{(h-1)}\right) = \frac{\pi_{lk}^{(h-1)} \left\{ \prod_{s=1}^{S} \prod_{u=1}^{U_s} \left(\pi_{lksu}^{(h-1)}\right)^{I_{jis}^{u*}} \right\}}{\sum_{v=1}^{V} \pi_{lv}^{(h-1)} \left\{ \prod_{s=1}^{S} \prod_{u=1}^{U_s} \left(\pi_{lvsu}^{(h-1)}\right)^{I_{jis}^{u*}} \right\}}$$

3– Draw:

$$\left(\pi_W^{(h)} \mid W^{(h)}, \alpha_W\right) \sim \text{Dir}\left(\alpha_1 + \sum_{j=1}^{J} I\left(W_j^{(h)} = 1\right), \ldots, \alpha_L + \sum_{j=1}^{J} I\left(W_j^{(h)} = L\right)\right)$$

Where $I(w_j^{(h)} = l) = 1$ if $w_j^{(h)} = l$ and 0 otherwise;

4– For $l = 1, \ldots, L$ draw:

$$\left(\pi_{lx}^{(h)} \mid W^{(h)} = l, X^{(h)}, \alpha_{lx}\right) \sim \text{Dir}\left(\alpha_{l1} + \sum_{j,i:W_j^{(h)}=l} I\left(X_{ji}^{(h)} = 1\right), \ldots, \alpha_{lk} + \sum_{j,i:W_j^{(h)}=l} I\left(X_{ji}^{(h)} = K\right)\right)$$

Where $I\left(X_{ji}^{(h)} = k\right) = 1$ if $X_{ji}^{(h)} = k$ and 0 otherwise;

5– For $l = 1, \ldots, L$ and $t = 1, \ldots, T$ draw:

$$\left(\pi_{lt} \mid W^{(h)} = l, \mathbf{Z}_t^{obs}, \alpha_{lt}\right) \sim \text{Dir}\left(\alpha_{lt1} + \sum_{j:W_j^{(h)}=l} I_{jt}^{1*}, \ldots, \alpha_{ltR_t} + \sum_{j:W_j^{(h)}=l} I_{jt}^{R_{t^*}}\right);$$

6– For $l = 1, \ldots, L$ and $k = 1, \ldots, K$ and $s = 1, \ldots, S$ draw:

$$\left(\pi_{lks} \mid W^{(h)} = l, X^{(h)} = k, \mathbf{Y}_s^{obs}, \alpha_{lks}\right) \sim$$
$$\text{Dir}\left(\alpha_{lks1} + \sum_{j,i:W_j^{(h)}=l \cap X_{ji}^{(h)}=k} I_{jis}^{1*}, \ldots, \alpha_{lksU_s} + \sum_{j,i:W_j^{(h)}=l \cap X_{ji}^{(h)}=k} I_{jis}^{U_{s^*}}\right).$$

(B) Part 2. After I iterations:

7– (imputation pace) perform M draws from the distribution $\text{pr}(\pi \mid \mathbf{Y}^{\text{obs}}, \mathbf{Z}^{\text{obs}})$ estimated in paces 1–6: in particular, the m-th draw ($m = 1, \ldots, M$) must include $w_j^{(m)}, x_{ji}^{(m)}, \pi_{lt}^{(m)}$ and $\pi_{lks}^{(m)}$, $\forall j, i, t, s \in \{\mathbf{Y}^{\text{mis}}, \mathbf{Z}^{\text{mis}}\}$, the missing part of the dataset. Perform the m-th imputation for the items at level 2 by drawing:

$$\left(Z_{jt} \mid W_j^{(m)} = l, \pi^{(m)}\right) \sim \text{Multinom}\left(\pi_{lt}^{(m)}\right),$$

and the m-th imputation for the items at level 1 by drawing:

$$\left(Y_{jis} \mid W_j^{(m)} = l, X_{ji}^{(m)} = k, \pi^{(m)}\right) \sim \text{Multinom}\left(\pi_{lks}^{(m)}\right),$$
$$\forall \; Z_{ji}, Y_{jis} \in \{\mathbf{Y}^{\text{mis}}, \mathbf{Z}^{\text{mis}}\}.$$

4. Findings

The data I use in this research consists of two nested levels, that is, lower-level data or students data are in higher level data or school and teacher data. The variables used in this research are as following.

Level 1 (students): Including 35 math questions:

$$X_1 : \text{is the first math question},$$
$$\vdots$$
$$X_{35} : \text{the 35th math question}.$$

IDSEX: the gender of the students.

The math questions include multiple choice, open-ended, and short answer questions, which were coded differently, so these 35 questions were all recoded so that if they answered correctly, they were assigned code 1, otherwise, code 0.

Level 2 (school and teacher questions): including 11 variables.

Z_1: Teacher teaching experience, Z_2: Teacher gender, Z_3: Teacher age, Z_4: Teacher education, Z_5: Teacher perception, Z_6: Teacher success rate, Z_7: Teacher expectation, Z_8: Access to computer and tablet for computing, Z_9: Does the class have a computer?, Z_{10}: Does the school have a computer? and Z_{11}: Teacher job satisfaction.

These 11 variables were also coded to 1 and 0. We consider the variable Y_0, which refers to students' success in mathematics coded to 1 and failure in mathematics coded to 0, as the response variable.

The logistic model for all 47 variables at both levels is as following:

$$\text{logitPr}(Y_{ji0} \mid \mathbf{X}_{ji}, \mathbf{Z}_j) = \beta_{j0} + \beta_1 X_{ji1} + \beta_2 X_{ji2} + \cdots + \beta_{35} X_{ji35} + \beta_{36} X_{\text{IDSEX}}$$

for level 1 and

$$\beta_{j0} = \beta_{00} + \gamma_1 Z_{j1} + \gamma_2 Z_{j2} + \cdots + \gamma_{11} Z_{j11} + u_{j0}$$

for level 2.

Using the significance level that is at p-values smaller than 0.05, we define the multilevel logistic model as following. The logit model is in figure 2.

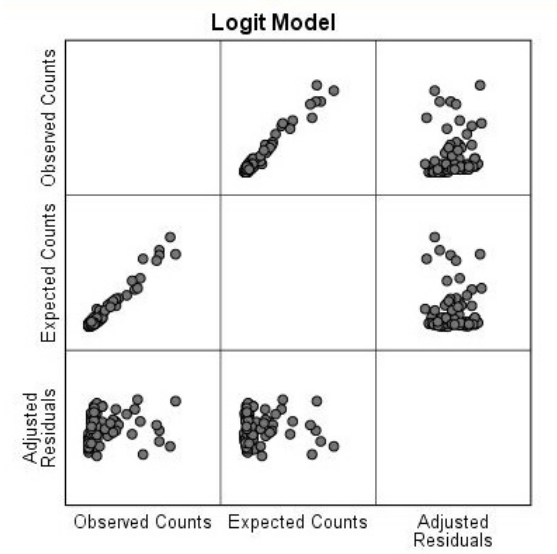

Figure 2: Logit model

$$\text{logitPr}(Y_{ji0} \mid \mathbf{X}_{ji}, \mathbf{Z}_j) = \beta_{j0} + \beta_{11} X_{ji11} + \beta_{19} X_{ji19} + \beta_{24} X_{ji24} \qquad (4.3)$$

for level 1 and

$$\beta_{j0} = \beta_{00} + \gamma_1 Z_{j1} + \gamma_6 Z_{j6} + u_{j0} \tag{4.4}$$

is for level 2. In Table 1, the values of all coefficients, that is, for 47 variables, are reported by BMLC and LD methods.

Table 1: Includes estimates, standard errors, and p-values obtained by LD and BMLC

Parameters	Estimates		Standard errors		p values		
	LD	BMLC	LD	BMLC	LD	BMLC	
β_{00}	$-1.30e^{+02}$	-1.92	$2.46e^{+05}$	0.58	1	0.00	$*$
β_1	$-2.77e^{+00}$	0.09	$1.40e^{+05}$	0.25	1	0.71	
β_2	$2.18e^{+01}$	0.33	$4.30e^{+05}$	0.34	1	0.32	
β_3	$2.85e^{+01}$	0.22	$1.3e^{+05}$	0.27	1	0.41	
β_4	$9.10e^{+00}$	0.48	$2.03e^{+05}$	0.27	1	0.07	
β_5	$-5.56e^{+00}$	0.05	$2.23e^{+05}$	0.29	1	0.85	
β_6	$-7.74e^{+00}$	0.14	$3.82e^{+05}$	0.35	1	0.67	
β_7	$-2.95e^{+01}$	-0.46	$1.61e^{+06}$	0.39	1	0.23	
β_8	$-1.47e^{+01}$	-0.51	$4.54e^{+05}$	0.41	1	0.22	
β_9	$1.93e^{+01}$	-0.07	$1.66e^{+05}$	0.28	1	0.79	
β_{10}	$4.44e^{+01}$	0.00	$2.39e^{+05}$	0.27	1	0.99	
β_{11}	$2.70e^{+01}$	0.59	$2.80e^{+05}$	0.29	1	0.04	$*$
β_{12}	$-5.29e^{+00}$	0.38	$1.01e^{+06}$	0.50	1	0.45	
β_{13}	$-1.17e^{+01}$	0.30	$3.7e^{+05}$	0.28	1	0.28	
β_{14}	$-4.38e^{+01}$	0.47	$2.50e^{+05}$	0.29	1	0.10	
β_{15}	$-6.10e^{-01}$	0.50	$1.94e^{+05}$	0.28	1	0.07	
β_{16}	$-1.72e^{+01}$	0.39	$3.18e^{+05}$	0.34	1	0.26	
β_{17}	$7.56e^{+00}$	0.34	$2.71e^{+05}$	0.32	1	0.29	
β_{18}	$1.44e^{+01}$	-0.52	$2.58e^{+05}$	0.38	1	0.17	
β_{19}	$2.74e^{+00}$	-0.88	$3.47e^{+05}$	0.40	1	0.02	$*$
β_{20}	$6.24e^{+01}$	0.26	$3.37e^{+05}$	0.36	1	0.46	
β_{21}	$-1.55e^{+01}$	0.00	$3.27e^{+05}$	0.39	1	0.99	
β_{22}	$-6.67e^{+00}$	0.99	$3.89e^{+05}$	0.52	1	0.05	
β_{23}	$1.13e^{+01}$	-0.06	$7.18e^{+05}$	0.27	1	0.81	
β_{24}	$3.93e^{+01}$	0.85	$1.91e^{+05}$	0.30	1	0.00	$*$
β_{25}	$-1.72e^{+01}$	0.04	$3.25e^{+05}$	0.34	1	0.90	
β_{26}	$-3.36e^{+00}$	0.69	$5.76e^{+05}$	0.45	1	0.12	
β_{27}	$8.64e^{-02}$	0.43	$6.74e^{+05}$	0.39	1	0.27	
β_{28}	$3.09e^{+01}$	0.03	$6.74e^{+05}$	0.29	1	0.91	

β_{29}	$1.89e^{+01}$	0.46	$3.82e^{+05}$	0.27	1	0.90	
β_{30}	$1.44e^{+01}$	0.36	$1.51e^{+05}$	0.26	1	0.16	
β_{31}	$-6.13e^{+00}$	−0.41	$1.31e^{+05}$	0.31	1	0.18	
β_{32}	$9.09e^{+00}$	−0.40	$3.12e^{+05}$	0.39	1	0.30	
β_{33}	$1.67e^{+01}$	−0.38	$4.31e^{+05}$	0.28	1	0.17	
β_{34}	$-3.21e^{+00}$	−0.16	$2.42e^{+05}$	0.26	1	0.52	
β_{35}	$1.49e^{+01}$	0.36	$3.03e^{+05}$	0.30	1	0.23	
β_{36}	$2.47e^{+00}$	0.06	$1.31e^{+05}$	0.26	1	0.80	
γ_1	0.62	0.57	0.28	0.28	0.02	0.04	
γ_2	−0.03	0.00	0.20	0.20	0.84	0.99	
γ_3	−0.16	−0.16	0.27	0.28	0.54	0.55	
γ_4	−0.03	0.02	0.26	0.27	0.89	0.92	
γ_5	−0.44	−1.58	0.72	1.14	0.53	0.16	
γ_6	1.85	2.64	0.84	1.16	0.02	0.02	*
γ_7	−0.17	−0.14	0.42	0.43	0.67	0.74	
γ_8	0.61	0.95	0.49	0.59	0.23	0.10	
γ_9	0.74	0.63	0.62	0.63	0.23	0.31	
γ_{10}	0.14	−0.22	0.51	0.60	0.77	0.70	
γ_{11}	0.17	0.32	0.41	0.46	0.67	0.48	
γ_{00}	−2.03	−1.86	0.94	1.01	0.03	0.06	

In the boxplot 3, in the horizontal part of the graph, which is the z-sum, the numbers 3, 4, 5, 6, 7, 8 refer to the number of features. The number 3 means that if we have 3 Z features, our prediction success is 4; if we have 4 features, our prediction success is 6, and the black line inside the rectangles indicates the average of our prediction. For 5 features, we have the number 24; for 6 features, the number 23; for 7 features, the number 9; and for 8 features, the number 17 in the boxplot. I had a limitation in drawing this graph, and no more than 8 cases were run. As you can see in Figure 3, the smaller the rectangle, the smaller the variance, and the larger the rectangle, the larger the variance in the figures. The top and bottom of each rectangle indicate our prediction confidence interval. For example, for 5 features, the confidence interval is $[11, 27]$; for 8 features, the confidence interval is $[3, 17]$; and so on.

5. Implementation of the methods

In the elementary stage, I implemented the BMLC model for Gibbs sampling (with 1000 iterations and 3000 estimation iterations) to obtain a posterior distribution of L and K for each incomplete dataset. I assumed $L = 20$ and $K = 20$ and

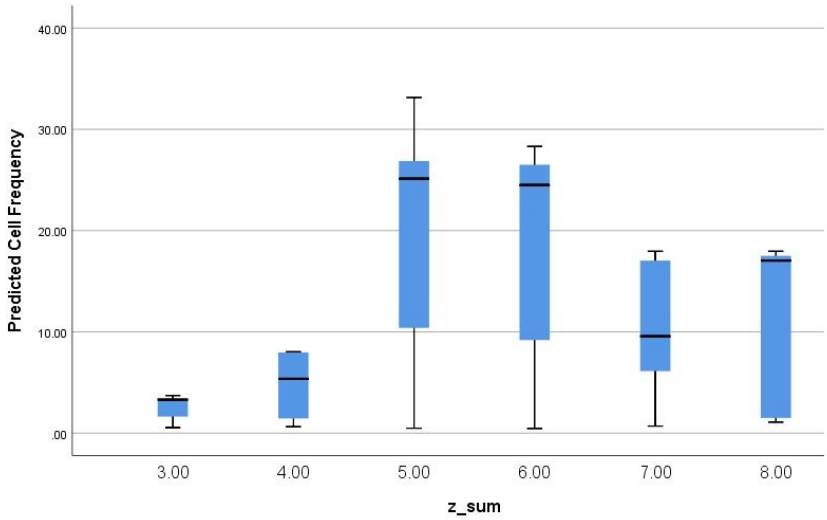

Figure 3: The boxplot

ran the program as in Algorithm 1, where L refers to the number of W_j, which is the number of latent classes at level 2, and K refers to the number of $X_{ji} \mid W_j$, which is the number of latent classes at level 1. After calculating the different iterations of the program that were performed in different paces, the optimal or maximum number of $L = 3$ and $K = 15$ for running the imputations was finally obtained. I considered the set of prior hyperparameters $\alpha_{ltr} = \alpha_{lksu} = 0.01$ for each l, k, t, s, r, u and the prior hyperparameters to ensure complete allocation $\alpha_l = 500$ for each l at level 2 and $\alpha_{lk} = 50$ for each l, k at level 1. $M = 5$ imputations were run, and the total Gibbs sampling iterations $I = 5000$, where $b = 2000$ for the burn-in and $I - b = 3000$ for the imputations, were run. All of this work was done with the R software as well as BMLC coding (or R coding), which is included in the appendix of Ghaffari Irdmousa (2020) and also in https://github.com/davidevdt/BMLCimpute.

6. Conclusion

The findings related to the TIMSS math test block 1 data were run with the LD and BMLC methods, and these two methods are reported in Table 1 for comparison based on estimates, standard errors, and p-values. In the second column of the estimates, in the fourth column of the standard errors, and in the sixth column, the p-values are reported. For decision-making, the null hypothesis coefficients are given, and those marked with $*$ are statistically significant, and in logistic

models 4.3 and 4.4, the models are written based on the parameters that are significant. The standard errors obtained with the LD method are larger than those reported for imputations with the BMLC method. On the other hand, with the BMLC imputation model, the full sample size can be exploited, and it also recovers the standard errors close to the full data cases. Bayesian estimators are obtained with the maximum posterior probability; moreover, a better classification is accomplished.

7. Discussion

In this article, given the importance of the non-responses problem in the data, which reduces the efficiency of estimators and often causes their bias, and considering the findings, in the present research, an attempt was made to examine the bias adjustment of missing data in the nested categorical dataset of the TIMSS test with multiple imputation. Accordingly, I propose the use of BMLC models for the MI of multilevel categorical data. After introducing the model, I performed a real study in order to assess its performance. The proposed BMLC model, a flexible imputation technique, is able to distinguish complex orders of relationships among variables in the dataset at both levels, returning unbiased and stable parameter estimates of the analysis model. This imputation model can be a useful tool for researchers dealing with missing multilevel categorical data. The model can be extended to data with three or more levels, for example, level 1 students, level 2 schools, and level 3 countries. Nevertheless, to approve our intuition, more inclusive research with different settings for the number of higher-level units and LCs, additionally for the value of the Level 2 mixture weights hyperparameter α_l, should be performed in the future.

Some suggestions that researchers should consider are:

1– To analyze data that contain missingness, make sure to use imputation so that non-response does not reduce the efficiency of the estimators and does not often bias them.
2– With large Level 1 data and small Level 2 data, run the BMLC model on the data and vice versa.
3– With longitudinal data, run the BMLC model.
4– With data that has a larger number of levels, run this model.

References

Ahmadi-Kahjouq, M. (2014). Study of missing data using Bayesian classification. Master's thesis, Allameh Tabataba'i University. [In Persian].

Allison, P. (2009). Missing data. In Millsap, R. E. and Maydeu-Olivares, A., editors, *The SAGE Handbook of Quantitative Methods in Psychology*, volume 4, page e6624. SAGE Publications.

Di Zio, M., Scanu, M., Coppola, L., Luzi, O., and Ponti, A. (2004). Bayesian networks for imputation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **167**, 309–322.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC Press.

Ghaffari Irdmousa, Z. (2014). Differential item functioning of mathematics test of TIMSS 2011 in grade 8 between girls and boys using the item response theory approach. Master's thesis, Islamic Azad University, Central Tehran Branch. [In Persian].

Ghaffari Irdmousa, Z. (2020). Bayesian analysis of missing data and its application in categorical data of the timss test. Master's thesis, Allameh Tabataba'i University, Faculty of Statistics. [In Persian].

Gilks, W. R. and Roberts, G. O. (1996). Strategies for improving MCMC. In *Markov Chain Monte Carlo In Practice*, volume 6, page 89–114. Springer.

Karimlou, M. and et al. (2006). Comparison of Bayesian and classical methods in estimating parameters of the logistic regression model with missing values in auxiliary variables. *Journal of the School of Public Health and Institute of Health Research*, **4**, 21–23. [In Persian].

Rashidi-Nejad, A. and Navabpour, H. (2010). Comparison of algorithmic imputation with two methods of mean imputation and new samples in panel surveys. *Iranian Official Statistics Review*, **21**, 72–89. [In Persian].

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, **63**, 581–592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, volume 81. John Wiley & Sons.

Van Buuren, S. and et al. (2011). Multiple imputation of multilevel data. In *Handbook of Advanced Multilevel Analysis*, volume 10, page 173–196. Springer.

Vermunt , J. K . (2003) . Multilevel Latent Class Models. *Sociological Methodology* , **33** , 213–239 .

Vermunt, J. K., Van Ginkel, J. R., Van der Ark, L. A., and Sijtsma, K. (2008). Multiple imputation of incomplete categorical data using latent class analysis. *Sociological Methodology*, **38**, 369–397.

Vidotto, D., Vermunt, J. K., and van Deun, K. (2018). Bayesian multilevel latent class models for the multiple imputation of nested categorical data. *Journal of Educational and Behavioral Statistics*, **43**, 511–539.

Vidotto, D., Vermunt, J. K., and Van Deun, K. (2020). Multiple imputation of longitudinal categorical data through bayesian mixture latent Markov models. *Journal of Applied Statistics*, **47**(10), 1720–1738.