International Journal of Assessment Tools in Education
2018, Vol. 5, No. 3, 403–416

DOI: 10.21449/ijate.430720

Published at http://www.ijate.net http://dergipark.gov.tr/ijate Research Article

Evaluating Performance of Missing Data Imputation Methods in IRT Analyses

Ömür Kaya Kalkan [1,*], **Yusuf Kara** [2], **Hülya Kelecioğlu** [3]

[1] Pamukkale University, Education Faculty, Measurement and Evaluation Department, 20160, Denizli, Turkey
[2] Anadolu University, Education Faculty, Measurement and Evaluation Department, 26470, Eskişehir, Turkey
[3] Hacettepe University, Education Faculty, Measurement and Evaluation Department, 06800, Ankara, Turkey

Abstract: Missing data is a common problem in datasets that are obtained by administration of educational and psychological tests. It is widely known that existence of missing observations in data can lead to serious problems such as biased parameter estimates and inflation of standard errors. Most of the missing data imputation methods are focused on datasets containing continuous variables. However, it is very common to work with datasets that are made of dichotomous responses of individuals to a set of test items to which IRT models are fitted. This study compared the performances of missing data imputation methods that are IRT model-based imputation (MBI), Expectation-Maximization (EM), Multiple Imputation (MI), and Regression Imputation (RI). Parameter recoveries were evaluated by repetitive analyses that were conducted on samples that were drawn from an empirical large-scale dataset. Results showed that MBI outperformed other imputation methods in recovering item difficulty and mean of the ability parameters, especially with higher sample sizes. However, MI produced the best results in recovery of item discrimination parameters.

ARTICLE HISTORY

Received: 10 May 2018
Accepted: 04 June 2018

KEYWORDS

Missing Data,
IRT Model-Based Imputation
Multiple Imputation,
Expectation-Maximization,
Regression Imputation

1. INTRODUCTION

Majority of research data in psychology, sociology, and education are collected from human subjects. Missing data is one of the most important problems for researchers working in these fields. Missing data generally occurs in situations where participants do not respond to some items in data collection process due to lack of knowledge, hesitation, and lack of motivation or planned missing data designs, etc. (Enders, 2010; Finch, 2010; Graham, Taylor, Olchowski, & Cumsille, 2006; Sijtsma & van der Ark, 2003; Little & Rubin, 2002). It is difficult to conduct statistical analyses and calculate total scale scores in the presence of missing data. Moreover, existence of missing observations in data results in problems such as biased parameter estimates, inflation of standard errors, loss of information, and weak generalizability of results (De Leeuw, Hox, & Huisman, 2003; Dong & Peng, 2013; Finch, 2010; Rubin 1987; Schafer, 1997). While determining a proper method for handling missing observations, researchers need to address the rate, mechanism, and pattern of missing data

CONTACT: Ömür Kaya Kalkan ✉ kayakalkan@gmail.com Pamukkale University, Education Faculty, Measurement and Evaluation Department, 20160, Denizli, Turkey

ISSN-e: 2148-7456 /© IJATE 2018

(Dong & Peng, 2013; Enders, 2010; Little & Rubin, 2002; Rubin, 1987; Schafer, 1997; Schafer & Graham, 2002).

1.1. Rate of Missing Data

Rate of missing data is the first issue that needs to be taken into account. Although there is not a certain criterion in the literature with regard to an acceptable rate of missing data in order to obtain valid inferences from a dataset, Tabachnick and Fidell (2007) and Schafer (1999) suggest that 5% or lower rate of missing data in a large sample would be inconsequential. On the other hand, Bennet (2001) suggests that parameter estimates through a statistical analysis are more likely to be biased, when more than 10% of data is missing. After addressing the rate of missing data, another issue that needs to be considered is the mechanism, namely how the missing data are distributed.

1.2. Missing Data Mechanism

Mechanism of missing data in a dataset can be Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR) (Little & Rubin, 2002; Rubin, 1976). In MCAR case, missing data do not follow a specific pattern and variables containing missing observations do not depend on any other observed variables. Although Little (1988) proposed a test for detection of MCAR data mechanism, in practice, it is quite difficult to determine whether missing data are MCAR or not (Little & Schenker, 1995; Schlomer, Bauman, & Card, 2010). MCAR can also be regarded as a special case of MAR or a random sample of a complete dataset (Schafer & Graham, 2002). While omitting MCAR data does not lead to bias, standard errors of the point estimates increase since the sample size decreases (Dong & Peng, 2013). In case of MAR mechanism, missing observations in a variable are related to other variables in a dataset, rather than the latent level of the measured variable itself (Allison, 2001; De Leeuw, Hox, & Huisman, 2003; Enders, 2004, 2010). As a result, MCAR and MAR do not follow a specific pattern and missing observations are rather a result of randomness. In MNAR, however, missing data have a specific pattern and cannot be easily omitted because they are associated with the latent levels of the measured variable (Enders, 2004; Tabachnick & Fidell, 2007). Besides the aforementioned mechanisms, there are also planned missing data designs that are widely preferred by practitioners, especially in large-scale assessment studies. In such data collection designs, not all items are responded by all individuals, due to time and cost limitations. Interested readers are encouraged to see Enders (2010) and Graham et al. (2006) for more detailed information about planned missing data.

1.3. Patterns of Missing Data

Enders (2010) stated that missing data pattern simplistically explains position of the "holes" in a data matrix. There are three types of missing data patterns that are referred to as univariate, monotone, and arbitrary, respectively. If responses of individuals to a set of items have missing observations related to one or more variable(s), this indicates a univariate pattern of missing data. Monotone missing data pattern is often observed in longitudinal studies, where data are collected repeatedly from same individuals. If missing data are randomly observed for any individual in any of the variables, this indicates an arbitrary missing data pattern. Computation of a univariate or monotone missing data pattern is easier than an arbitrary pattern (Dong & Peng, 2013; Little & Rubin, 2002). Below, widely-used methods for handling various types of missing data are briefly explained.

1.4. Methods for Handling Missing Data

According to Little and Rubin (2002), most of the missing data handling methods can be classified into four general categories as follows: procedures based on completely recorded units, weighting procedures, imputation-based procedures, and model-based procedures.

Primitive methods like List-wise Deletion (LD) and Pair-wise Deletion (PD) simply suggest deletion of cases that have missing observations. LD, which proposes deletion of all cases that contain missing values for any variable, is not a recommended method in general, because it leads to loss of statistical power. It is also known that PD, which excludes cases that have missing values for variables covered, causes problems in correlation based calculations and multivariate analyses (Schlomer, Bauman, & Card, 2010). In sum, these two methods, which are based on exclusion of cases that have missing data, often lead to biased and inefficient parameter estimates (Rubin, 1987; Schafer, 1997).

Apart from deletion-based solutions, several methods have been developed for imputing the missing values by various techniques. Mean Substitution (MS) is one of the simplest imputation methods, where missing values are replaced with mean of the existing observations of the relevant variable. Although it sounds convenient to implement, MS results with reduced variance, which in turn leads to biased parameter estimates (Allison, 2001; Bennett, 2001; Graham, Cumsille, & Elek-Fisk, 2003; Pallant, 2007). Another relatively primitive method, which is called as Regression Imputation (RI), uses the variable(s) that does not contain any missing values as predictors in a regression model, in order to impute missing observations. Although it yields unbiased mean values in MCAR and MAR, RI leads to bias in variance and covariance (Graham, Cumsille, & Elek-Fisk, 2003).

There are more sophisticated imputation methods that have a wide range of usage including Multiple Imputation (MI), Expectation-Maximization (EM), Pattern-Matching Imputation (PMI), Hot-deck Imputation (HDI) and many more. In addition to RI, we only included MI and EM among more sophisticated imputation methods in this study. Thus, a brief information is provided below for MI and EM, as well as another method that is called as Model Based Imputation (MBI). For a detailed overview of the all mentioned imputation methods, interested readers can see Finch (2010), Little and Rubin (2002), Schafer and Graham (2002), Schlomer, Bauman, and Card (2010).

1.5. MI and EM

Rather than replacing missing observations directly, MI and EM combine the existing information obtained from data that meet some statistical assumptions to estimate the missing data mechanism. MI first creates a pre-specified number of copies of a dataset, in which missing observations are appropriately imputed. In other words, MI produces multiple datasets that have different imputed observations for missing cells. Then, parameters are estimated from each of the imputed datasets based on the model to be fitted. In the final step, multiply estimated parameters are averaged to provide a single estimate (Royston, 2004).

EM is an iterative procedure, where initial estimates of the missing values are imputed by a regression model that includes a random error term. In the following step, the covariance matrix and set of means is estimated. Then, previously obtained covariance matrix and means are used in another regression model for the estimation of missing observations. These steps (i.e., Expectation [E] and Maximization [M]) continue until the change in the covariance matrix reaches a minimum value. When this condition is satisfied, the iteration stops and final imputations are used for the desired analysis (Allison, 2001; Finch, 2008). It is important to note that with categorical data analysis, it is necessary to round these imputed values, as they will typically not be integers.

1.6. MBI

Briefly stated, MBI imputes the missing values based on statistical model that is applied to the data. There are various MBI methods that can be used for imputation of missing data in IRT applications. In one of them, Mellenbergh (2002) proposed a procedure based on computation of a π parameter by the response pattern, score, and frequency of the observed

data. Missing responses then imputed by randomly sampling an observation from a Bernoulli distribution with that parameter. Another method based on non-parametric IRT models proposed by Sijtsma and van der Ark (2003), and called as response function imputation (RFI). RFI requires computation of a fraction that is based on total score, summary score, and rest score (which is obtained by omitting the selected item from the total test score) for the imputation of missing observations. For a more detailed information, interested readers can see the referred studies.

MBI that is adopted in this study is rather a simpler procedure compared to mentioned approaches. Strictly speaking, we employed 3 parameter IRT model's equation for imputing missing observations by using the information from the complete data (i.e., dataset with no missing observations). Details about this procedure are elaborated in method section.

1.7. Purpose of the Current Study

A review of the literature reveals that the missing data itself and missing data handling methods in different sample sizes have a direct effect on the IRT model applications (Enders, 2004; Finch, 2008; Glas & Pimentel, 2008; Huisman, & Molenaar, 2001; Sijtsma & van der Ark, 2003). In addition, Dong and Peng (2013) underlie the fact that majority of the studies that have been conducted in the last decade do not contain sufficient information about missing data. Namely, they have not defined a certain approach in handling missing data, and have not tested assumptions about missing data methods.

Besides the limited reports about procedures of handling missing data issues in social science researches, majority of the proposed methods focus on continuous data. However, binary response patterns obtained from achievement tests that are administered to a group of students are one of the most common types of categorical data in educational sciences. Moreover, most of the widely used psychological tests are made of items that have multiple category response options (e.g., likert-type structured response scales). Common procedure for handling missing observations in such datasets is using the mentioned methods that are developed for continuous missing observations. Appropriateness of these imputation techniques for categorical missing data has not been clearly expressed in the literature. This study primarily aims to fill this gap by inspecting the performance of these methods in datasets that consist of binary response patterns to which IRT models are fitted. In addition to widely-used EM, MI, and RI methods, MBI was also evaluated in the study. Thus, the performance of the mentioned methods under different amounts of MCAR data were compared based on the recovery of item and ability parameter estimates. On the other hand, it is worth noting that there is a limited number of studies that use MBI for imputing dichotomous missing observations in datasets to be analyzed by IRT models. Therefore, we believe that this study provides significant contributions both to the existing literature and practitioners about handling missing data in IRT modeling applications.

2. METHOD

2.1. Data Set

Research data consist of binary (i.e., 0-1) coded responses of 480,691 candidates to 19 multiple choice items in the Turkish language test, which is a subtest of the 6th Grade Placement Exam (SBS) conducted by the Ministry of National Education of the Republic of Turkey in 2008. SBS is an annual nation-wide exam that measures achievement levels of 6th, 7th and 8th graders in Turkish, Maths, Science and Technology, and Social Science courses. Although 958,879 candidates took the exam in 2008, only a subset of 480,691 candidate's data were obtained with official permission. This raw dataset contained many missing observations, as was expected. Thus, a clean dataset of 306,757 individuals who answered all items (namely,

the dataset that didn't have any missing observations) was drawn from the mentioned raw dataset. We named this cleaned dataset as "full dataset" in order to prevent any conflicts with the raw dataset that includes missing cells. In that case, the full dataset can be realized as the population data for the sake of our recovery evaluation study. In other words, the "true" values of the item and ability parameters are estimates that would be obtained by fitting the relevant IRT model to the "full dataset", namely the population dataset.

Obtaining random samples of complete data sets. As a second step of data preparation, samples of 250, 500, 1,000, and 5,000 (25 replications for each sample size) were drawn repeatedly from the full dataset by simple random sampling employing SPSS (IBM, 2011). In order to determine the best fitting IRT model, -2LL values that were obtained by applying unidimensional 1PL, 2PL, and 3PL models to the full dataset were evaluated. The 3PL model was found to be the one that fits best among the three models. Therefore, unidimensional 3PL model was used in all subsequent analyses with the random samples that were drawn from the full dataset.

Obtaining datasets that contain missing observations. Only the missing data that would show MCAR were considered in this study. "missForest" (Stekhoven, 2016) package of R (R Core Team, 2016) was used to obtain datasets with MCAR data. Thereby, datasets with missing observations were obtained by deleting 5%, 10%, and 15% of the observations from complete datasets sampled in different sizes from the full dataset. Thus, 25 datasets with MCAR patterns were created for each of the 12 conditions (4 sample sizes x 3 missing data rates), leading to a total of 300 datasets.

2.2. Missing Data Imputations

The EM, MI, RI, and MBI based on unidimensional 3PL IRT model probability function were used to impute the missing observations. SPSS (IBM, 2011) was used to impute missing data through the EM, MI, and RI. Imputations with EM were performed based on normal distribution and the maximum number of iterations was 25 as default. This quantity specifies the number of iterations that is used to estimate the true covariance. When the specified number of iterations is reached, the procedure stops even if the estimates have not converged (IBM, 2014). Since the EM algorithm is an iterative process, decimal values are likely to be obtained in categorical data. Hence, all decimal values are rounded to nearest integers in order to end up with a categorical outcome (Finch, 2008).

Imputations based on MI were performed through the Markov chain Monte Carlo (MCMC) algorithm and maximum number of iterations was set to the default value of 10 as defined in SPSS (IBM, 2011). Maximum number of iterations taken by MCMC is used by fully conditional specification (FCS) method. FCS fits a univariate model using all other available variables in the model as predictors. The method then imputes missing values for the variable being fit for each iteration and each variable. This procedure continues until the maximum number of iterations is reached (IBM, 2014). Finally, FCS iteration history data are examined to evaluate the convergence of the model. For any missing observation in the missing datasets, as suggested by Royston and White (2011), the number of imputations was in accordance with the rate of missing data (5, 10, and 15). Thus, for a dataset containing 5% missing data, 5 different datasets were obtained by MI. Similarly, 10 and 15 different datasets were obtained for datasets containing 10% and 15% missing data, respectively. Each MI dataset for each missing rate (e.g., 5 MI datasets for 5% missing rate) was analyzed separately and results (e.g., item parameter estimates) were combined as

$$\bar{Q} = \sum_{m} \hat{Q}_m \, / \, N$$

where N is the number of unique analyses (Finch, 2008).

Imputation that was performed by RI was also accomplished using SPSS (IBM, 2011). RI in SPSS (IBM, 2011) uses multiple linear regression for estimating the missing observations (IBM, 2011, 2014). Residuals of that regression model are also included for adjustment of the estimates. Thus, error terms are drawn randomly from the observed residuals of complete cases and added to the regression estimates (IBM, 2014). Then, decimal values obtained for each missing observation are rounded to nearest integers, just as the final step of EM.

In imputations with MBI, 3PL IRT model was used because it provided the best model fit with the full dataset, as mentioned in the preceding section. The statistical equation of 3PL IRT model is shown below.

$$P(X_{is}=1|\theta_s,\beta_i,\alpha_i,\gamma_i)=\gamma_i+(1-\gamma_i)\frac{\exp[\alpha_i(\theta_s-\beta_i)]}{1+\exp[\alpha_i(\theta_s-\beta_i)]} \quad (1)$$

where;

X_{is}= response of person s to item i (0 or 1),

θ_s= trait level for person s,

β_i= difficulty of item i,

α_i= discrimination for item i,

γ_i= lower asymptote (guessing) for item i (Embretson & Reise 2000; Hambleton & Swaminathan 1985).

When imputing values by MBI, a two-step procedure is followed as follows: In the first step, datasets with missing observations that were crafted for each condition were analyzed based on 3PL model via BILOG controlled by the "irtoys" (Partchev, 2016) package of R (R Core Team, 2016). “NA” value was assigned to the missing cells of the datasets and these cells were handled as missing in the analyses (Partchev, 2016). In that case, item responses that were missing for an individual were not included in the estimation of item and ability parameters. Item and person parameter estimations were performed using ML and Marginal Maximum Likelihood (MML), respectively. In the second step, in order to impute the missing observations, the item and ability parameters obtained from the first step were used. Namely, for a missing response of an individual to an item, the probability was calculated by substituting the relevant item and ability parameters into equation 1. Then, if the calculated probability is equal to or greater than 0.5, “1” was placed into the missing item response for that individual. Otherwise, “0” was imputed for the relevant cell.

2.3. Evaluating the Accuracy of the Imputation Methods

A total of 48x25=1,200 analyses were performed with the 25 datasets crafted for each of the 48 conditions (4 sample sizes x 3 missing data rates x 4 imputation methods). Item and mean of the ability estimates were compared with the “true values” that were obtained from the population, namely from the “full dataset”. Population estimates of the item parameters that are considered as true values for recovery evaluation are provided in Table 1.

Table 1. Population estimates of item parameter and standard errors

	Parameter Estimate (SE)		
Item	a	b	c
1	3.312 (0.025)	-0.118 (0.004)	0.374 (0.002)
2	2.497 (0.018)	-0.320 (0.006)	0.304 (0.003)
3	1.209 (0.011)	-0.659 (0.019)	0.212 (0.008)
4	1.864 (0.012)	-0.823 (0.009)	0.150 (0.005)
5	2.038 (0.013)	-0.688 (0.007)	0.184 (0.004)
6	2.518 (0.015)	-0.666 (0.005)	0.170 (0.003)
7	2.528 (0.014)	-0.587 (0.005)	0.123 (0.003)
8	1.139 (0.006)	-1.170 (0.006)	0.003 (0.001)
9	1.585 (0.013)	0.480 (0.007)	0.205 (0.003)
10	1.904 (0.013)	-0.387 (0.007)	0.169 (0.003)
11	2.321 (0.016)	-1.467 (0.009)	0.117 (0.007)
12	1.889 (0.014)	-0.382 (0.008)	0.244 (0.004)
13	1.925 (0.015)	0.724 (0.005)	0.175 (0.002)
14	2.249 (0.016)	-0.638 (0.007)	0.297 (0.004)
15	1.627 (0.010)	-0.662 (0.009)	0.092 (0.005)
16	2.165 (0.014)	-0.311 (0.005)	0.184 (0.003)
17	2.690 (0.018)	-0.744 (0.006)	0.243 (0.004)
18	2.471 (0.022)	0.543 (0.005)	0.383 (0.002)
19	2.240 (0.015)	0.039 (0.005)	0.207 (0.002)

As can be inspected from Table 1, SE's of the population estimates are considerably lower than .05, which can be regarded as a sign of accurate estimation. Majority of the difficulty parameters are lower than 0, indicating a relatively easy test. Discrimination parameters are all higher than 1, which is a sign of good discriminating test for low and high ability groups of individuals. Guessing parameters on the other hand, varied between 0.003 and 0.383.

Due to varying number of the sample sizes for each condition, recovery of the ability parameters was evaluated based on their calculated means, which should be zero according to 3PL IRT models' identification constraint. Thus, we compared the mean of the estimated ability parameters from each condition by the mean of population estimates, which was calculated as 0.18 in the full dataset. Average of the absolute differences (AAD) among replications were calculated as a recovery indicator for ability parameters by the equation provided below.

$$AAD = \frac{\sum |\bar{\hat{\theta}}_m - 0.18|}{25} \quad (2)$$

$\bar{\hat{\theta}}_m$ is the mean of the estimated ability parameters at the *m*th step of the replications. Performance of the imputation methods based on recovery of the true item parameters were compared through RMSE (Root Mean Square Error) by the formula given below.

$$\text{RMSE} = \sqrt{\frac{\sum(\hat{\alpha}_m - \alpha)^2}{25}} \quad (3)$$

$\hat{\alpha}_m$ is the item parameter value estimated from imputed dataset at the *m*th step and α is the parameter value estimated from population, namely the so-called true value. RMSE for item parameters is regarded as indicator of the variability in the estimates. We reported the mean of the RMSE's for each type of item parameter. Thus, mean RMSE values for difficulty, discrimination, and guessing parameter, as well as, an AAD value for mean of the ability parameters were reported.

3. RESULTS

3.1. Results for Item Parameters

Recovery performance for discrimination, difficulty and guessing parameters are presented separately by tendency graphs of mean RMSE's below. Looking at the all figures for all item parameters, one can observe that mean RMSE's decreases for all methods by the increase of the sample size, with any rate of missing data. This is an expected improvement thanks to increase of available data that provides extra information for more accurate parameter estimations. On the contrary, as the rate of missing data increases, mean RMSE's for all methods also increases with any of the sample sizes. This phenomenon is also in line with the previous fact that we just mentioned. Namely, the amount of information lost increased due to increased number of missing observations, which in turn leads accuracy of the parameter estimates to decrease.

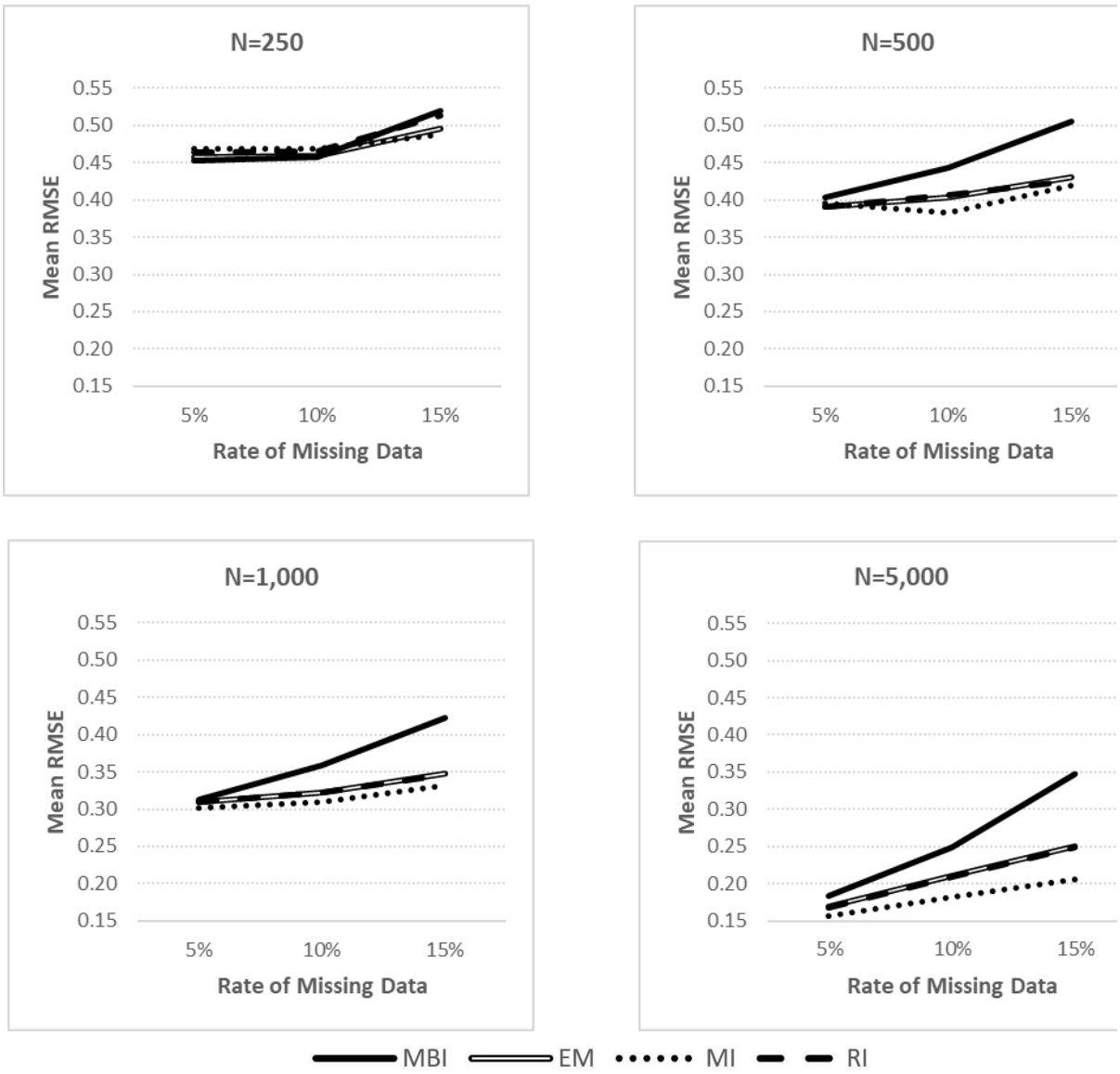

Figure 1. Mean RMSE values for item discrimination parameters

Looking at Figure 1 for discrimination recoveries, it can be observed that all imputation methods had nearly identical mean RMSE's when missing data rate is 5%, with all sample sizes. With N=250, all methods seemed to behave nearly in a similar fashion, in terms of mean

RMSE change. With all other sample sizes, EM and RI showed nearly undistinguishable trend of mean RMSE change. On the other hand, MBI apparently had higher mean RMSE's with all sample sizes (except for N=250) with all rates of missing data, while MI had slightly lower ones. Moreover, increase in mean RMSE was steeper for MBI, compared to all three methods, especially with three larger sample sizes.

Trend of the mean RMSE's for difficulty parameters are illustrated in Figure 2. Different from the previous findings for discriminations, MBI was the method that had the lowest mean RMSE's nearly in all conditions for item difficulty recoveries. Moreover, the increase in mean RMSE by the inflation of missing rates, was not as dramatic for MBI as other method's, especially with the two larger sample sizes.

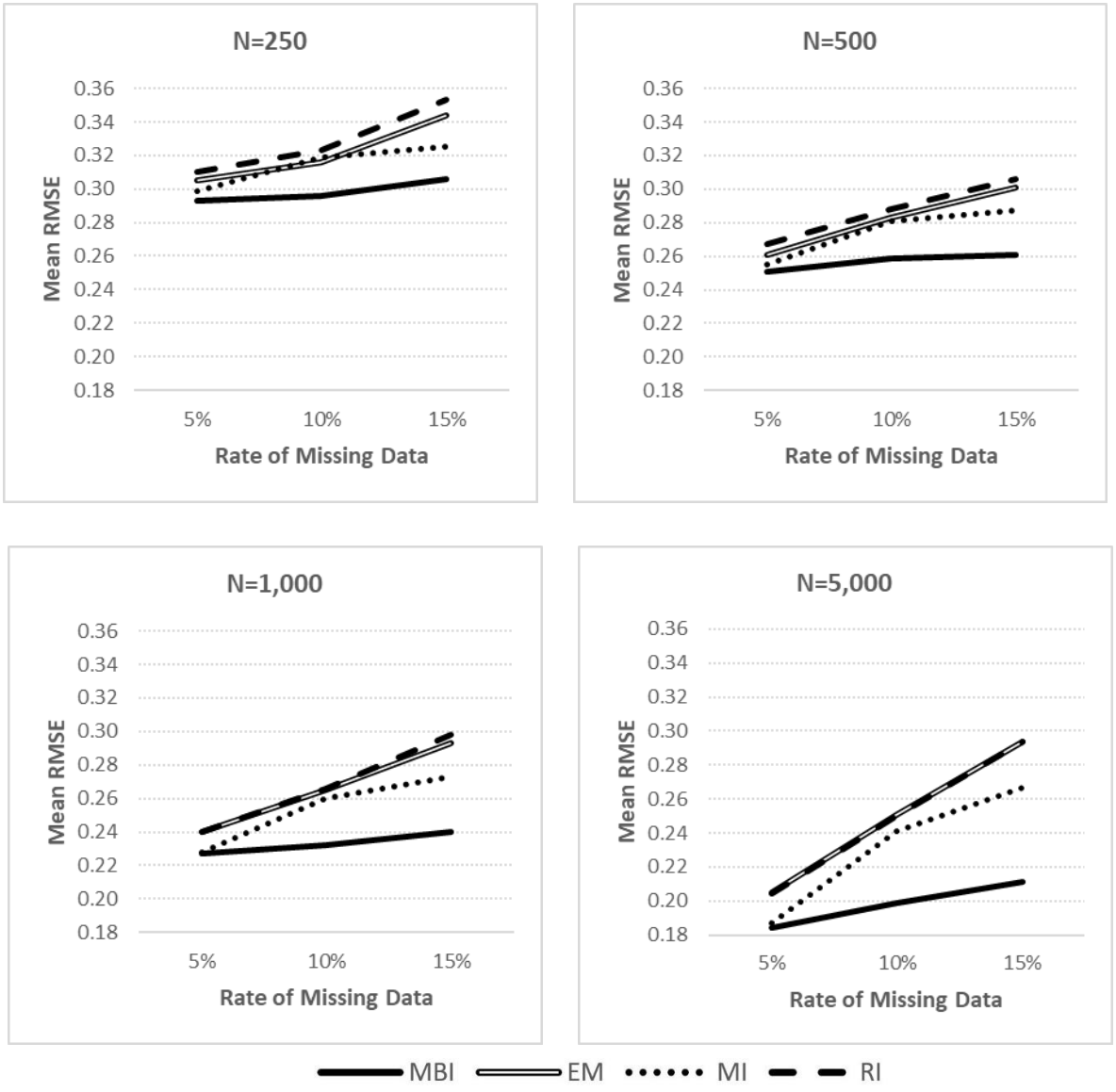

Figure 2. Mean RMSE values for item difficulty parameters

Similar to previous finding in discrimination parameters, EM and RI had nearly identical mean RMSE trend for the recovery of item difficulties.

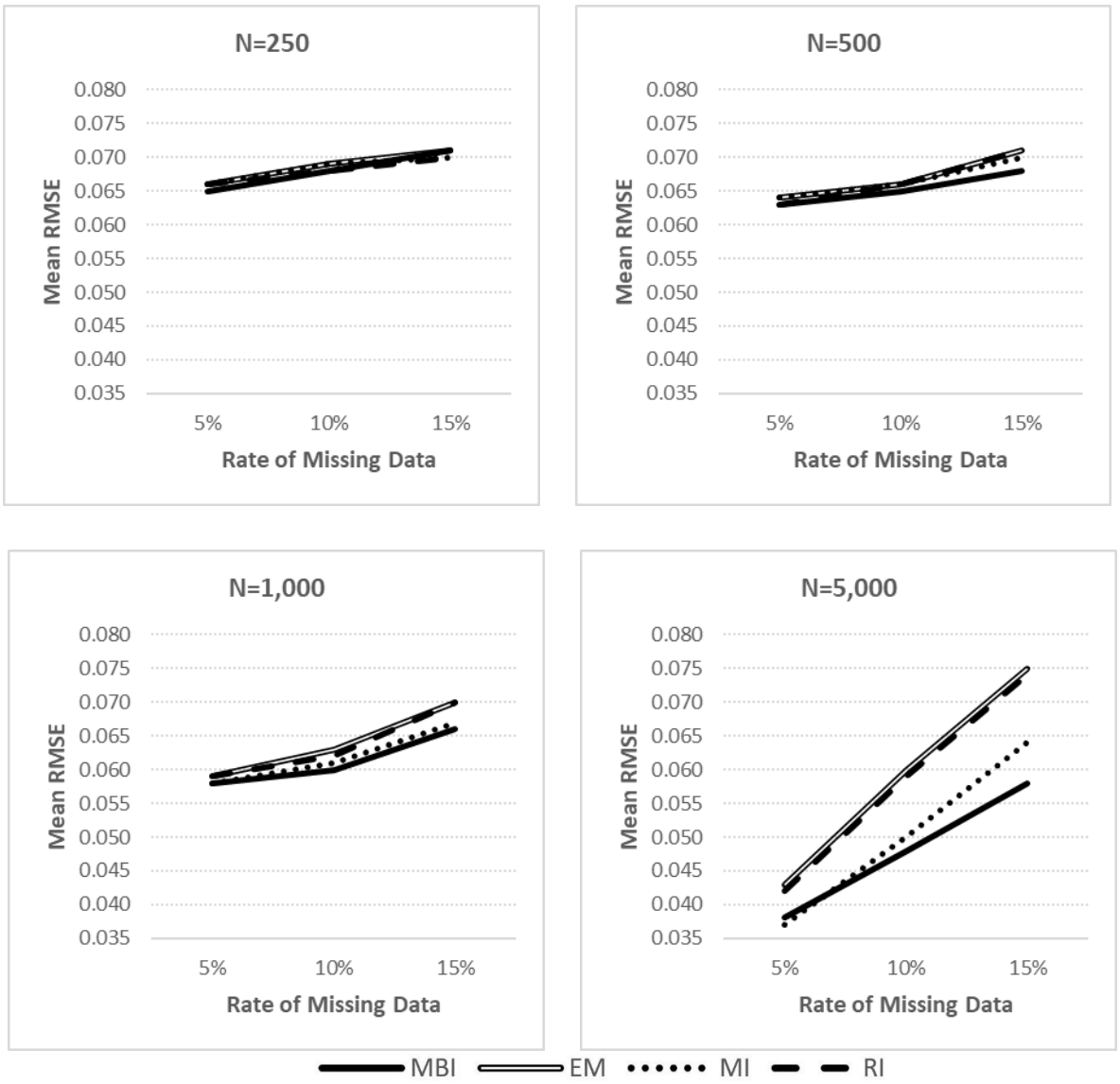

Figure 3. Mean RMSE values for guessing parameters

Trend of the mean RMSE's for guessing parameters are illustrated in Figure 3. Difference among methods is nearly indistinguishable for N=250, and it is quite similar for N=500 and N=1,000. For the largest sample size, differences among methods became more observable, except MI and RI, which behaved nearly identical. Moreover, MBI had relatively smaller mean RMSE's compared to other methods, especially with larger rates of missing data, when N=5,000.

3.2. Results for Ability Parameters

Graphs for the change of the averaged absolute differences (AAD) between means of the true and estimated ability parameters are presented in Figure 4. All methods had close AAD values with the lowest rate of missing data in all sample sizes. MBI had relatively smaller AADs with 10 and 15% rates of missing data. EM and RI again had nearly identical AAD trends, with N=1,000 and N=5,000. As a general inspection, it can be said that RI can be considered the method that had relatively higher AAD's compared to other ones. Last, it is also important to imply that the increase in AAD's was nearly linear when N=5,000 for all methods.

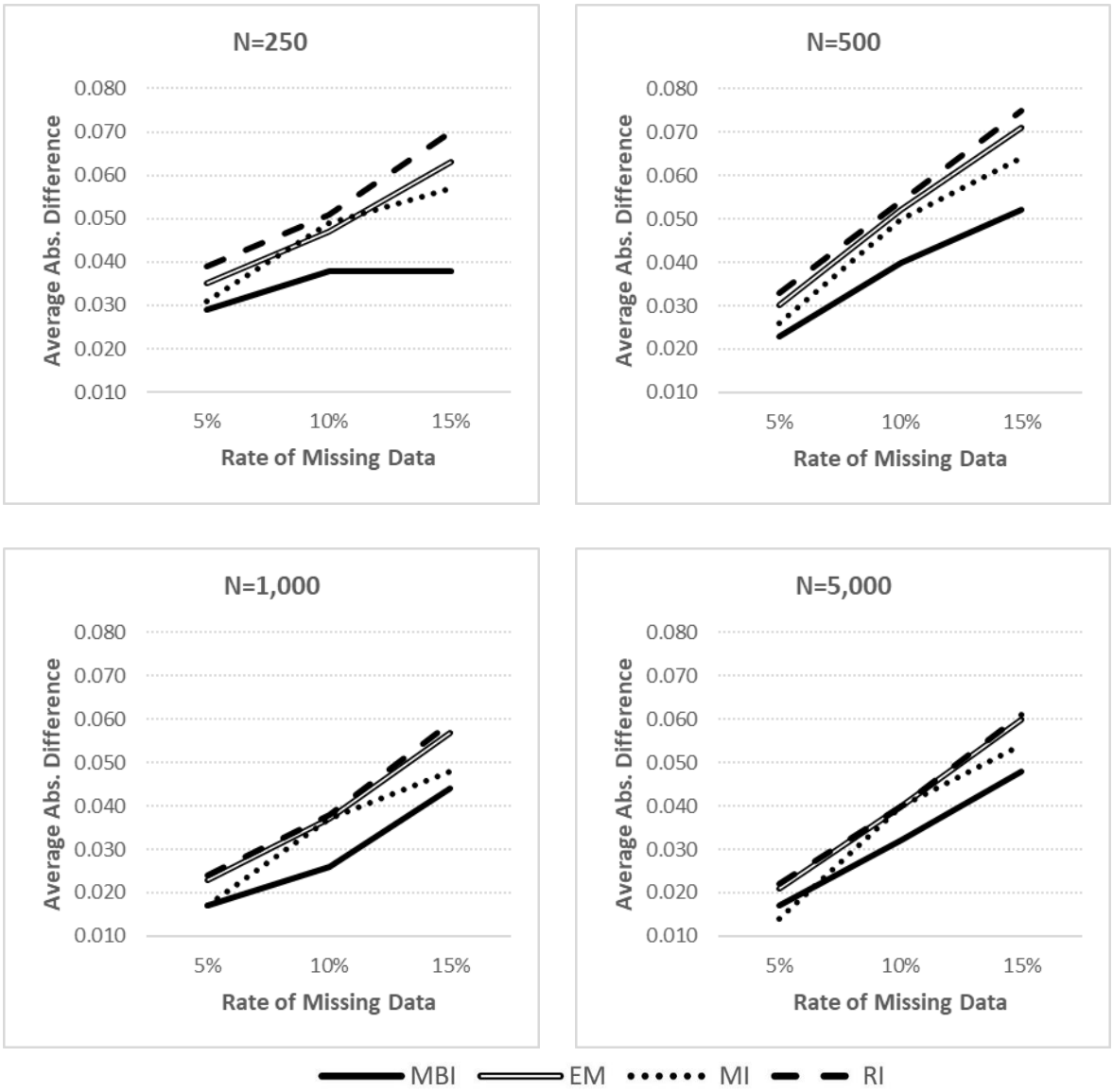

Figure 4. Average absolute difference values for the mean of the ability parameters

4. DISCUSSION AND CONCLUSION

In this study, the performance of various missing data imputation methods in terms of IRT models' parameter recovery were compared under different sample sizes and missing data rates. For the lowest rate of missing data (5%), it can be concluded that all methods showed close recovery performance for item and mean of the ability parameters in all sample sizes. Moreover, this similar performance occurred at lower levels of mean RMSE's for the two largest sample sizes. These conclusions in fact are parallel with Tabachnick and Fidell (2007). Namely, they stated that 5% or less MCAR data would not cause serious problems in large datasets. Moreover, they also stated that various missing data imputation methods are more likely to perform similarly in such conditions.

Methods showed different performances across different types of item parameters. Namely, MI generally was more robust to the increase of the missing data rate in recovery of the discrimination parameters. In fact, there are numerous studies which report that MI provides generally less-biased estimates compared to single-imputation methods. For example, Schlomer, Bauman, and Card (2010), and Schafer and Graham (2002) reported that the

accuracy of the parameter and standard error estimates makes MI the best choice for handling missing data. Similarly, it was also reported that MI produces unbiased parameter estimates in datasets that contain MAR data pattern (Peugh & Enders, 2004) and performs better than EM (Acock, 2005). Finch (2008) also verified MI's less-biased recovery performance in IRT model estimates, compared to other methods. MI's for discrimination parameters can also be attributed to performing imputation to each dataset according to the rate of missing data (Royston & White, 2011). EM and RI performed nearly identical and took the second place after MI in terms of recovery of the discrimination parameters, while MBI took the last place. It is important to note that these prominent implications were true especially for sample sizes 500 or larger, because all methods behaved quite similar with the smallest sample size N=250. As implied by Finch (2008), EM's relatively low recovery performance can be attributed to its assumption of normality, which is not the case for the dichotomous item responses.

When it comes to item difficulties, MBI showed the best recovery performance with all sample sizes that have missing data especially 10% and more. MI can be placed in the second order, while EM and RI again showed nearly identical behavior as the last two methods. For guessing parameters, all methods behaved nearly identical at all rates of missing data for the sample sizes of 250, 500 and 1,000. In fact, Finch (2008) reported that the estimation of the guessing parameters were not dramatically differed among the methods that he covered. The distinction among methods became more apparent with the maximum sample size of 5,000, especially for 10% or larger missing data. MBI again was slightly more robust to the increase of missing data, compared to other methods for the sample size of 5,000. MBI recovered the mean of the ability parameters better than other methods, with the 10% and 15% of missing data. For the 5% missingness, the distinction among methods was not that considerable.

Unfortunately, as also implied by other researchers (e.g., Finch, 2008; Smits, Mellenbergh, & Vorst, 2002) it is hard to recommend a single imputation method that will work in every scenario. Depending on our results, we can say that apart from the well-known methods that are MI, EM and RI, MBI can also be used as an alternative for dichotomous data containing missing observations in IRT model applications. MBI will particularly be more effective in scenarios where the ability and item difficulty estimations are more of interest. Thus, the use of MBI is expected to be more effective in Rasch and 1 PL models, where item discrimination parameters are not modeled. Nevertheless, if item discrimination parameters are of more interest, it would be more reasonable to prefer MI. Moreover, it is also important to imply that differences among the performances of MBI and MI was not so dramatic in many conditions. Thus, if two or three parameter models are desired and all model parameters would have the same level of interest for the practitioner, MI would be more rationale to adopt. Moreover, MBI requires long calculations for each of the missing observations. Considering this limitation of MBI, MI also offers a more reasonable solution in terms of efficiency.

There are some further research suggestions that we draw depending on the results of that study. First, evaluating the performance of MBI with one and two parameter IRT models and comparing the results with the current study is thought to be worth further investigation. In addition, the performance of MBI can also be examined in multidimensional datasets that contain missing data. Last, the consequences of assigning 0 to the missing observations (treating missing observations as false) rather than handling them as missing can also be explored for MBI.

Acknowledgements

We declare that a part of this study was presented as an oral presentation at the IIIrd International Eurasian Educational Research Congress (EJER, 2016) held on 31 May 03 June 2016 in Muğla, Turkey.

ORCID

Ömür Kaya Kalkan https://orcid.org/0000-0001-7088-4268
Hülya Kelecioğlu https://orcid.org/0000-0002-0741-9934

5. REFERENCES

Acock, A. C. (2005). Working with missing values. *Journal of Marriage and Family, 67(4),* 1012-1028.

Allison, P. D. (2001). *Missing data*. Thousand Oaks, *CA, US: Sage publications*.

Bennett, D. A. (2001). How can i deal with missing data in my study? *Australian and New Zealand journal of public health*, *25*(5), 464-469.

De Leeuw, E. D., Hox, J., & Husman, M. (2003). Prevention and treatment of item nonresponse. *Journal of Official Statistics, 19*(2), 153-176.

Dong, Y., & Peng, C. Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus*, *2*(1), 222.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Enders, C. K. (2004). The impact of missing data on sample reliability estimates: Implications for reliability reporting practices. *Educational and psychological measurement*, *64*(3), 419-436.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement*, *45*(3), 225-245.

Finch, W. H. (2010). Imputation methods for missing categorical questionnaire data: A comparison of approaches. *Journal of Data Science*, *8*(3), 361-378.

Glas, C. A., & Pimentel, J. L. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement*, *68*(6), 907-922.

Graham J.W., Cumsille, P.E., & Elek-Fisk, E. (2003). *Methods for handling missing data.* In Research Methods in Psychology, ed. JA Schinka, WF Velicer, pp. 87–114. Volume 2 of Handbook of Psychology, ed. IB Weiner. New York: Wiley

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological methods*, *11*(4), 323.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications* (Vol. 7). Boston, MA: Kluwer-Nijhoff Publishing.

Huisman, M., & Molenaar, I. W. (2001). Imputation of missing scale data with item response models. In *Essays on item response theory* (pp. 221-244). Springer, New York, NY.

IBM. (2011). IBM SPSS statistics base 20. *Chicago, IL: SPSS Inc*.

IBM (2011). IBM SPSS missing values 20. Retrieved January 24, 2018, from https://www.csun.edu/sites/default/files/missing-values20-64bit.pdf

IBM (2014). IBM SPSS missing values 22. Retrieved January 24, 2018, from http://www.sussex.ac.uk/its/pdfs/SPSS_Missing_Values_22.pdf

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83, 1198–1202.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. Wiley. New York.

Little, R. J., & Schenker, N. (1995). *Missing data*. In Handbook of statistical modeling for the social and behavioral sciences (pp. 39-75). Springer US.

Mellenbergh, G. J. (2002). *Measurement model-based imputation of missing item responses*. Unpublished manuscript.

Pallant, J. (2007). *SPSS survival manual* (3rd ed.). New York, NY: Open University Press.

Partchev, I. (2016). Irtoys: simple interface to the estimation and plotting of IRT models. *R package version 0.2.0*

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of educational research*, *74*(4), 525-556.

R Core Team. (2016). R: A language and environment for statistical computing [Computer Software]. Vienna, Austria: R Foundation for Statistical Computing.

Royston, P. (2004). Multiple imputation of missing values. *Stata journal*, *4*(3), 227-41.

Royston, P., & White, I. R. (2011). Multiple imputation by chained equations (MICE): implementation in Stata. *Journal of Statistical Software*, *45*(4), 1-20.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581-592.

Rubin, R. B. (1987). *Multiple imputation for nonresponse in surveys* (J Wiley & Sons, New York, NY).

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC press.

Schafer, J. L. (1999). Multiple imputation: a primer. *Stat Methods in Med 8*(1), 3–15.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, *7*(2), 147.

Schlomer, G. L., Bauman, S., & Card, N. A. (2010). Best practices for missing data management in counseling psychology. *Journal of Counseling psychology*, *57*(1), 1.

Sijtsma, K., & 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.

Smits, N., Mellenbergh, G. J., & Vorst, H. (2002). Alternative missing data techniques to grade point average: Imputing unavailable grades. *Journal of Educational Measurement*, *39*(3), 187-206.

Stekhoven, D. J. (2016). MissForest: nonparametric missing value imputation using random forest. *R package version 1.4.*

Tabachnick, B. G., & Fidell, L. S. (2007). *Using Multivariate Statisti*cs. Pearson Education. Boston, MA.