2018 ■ Vol. 14 ■ no. 3

Impact of simple substitution methods for missing data on Classical test theory difficulty and discrimination

Sébastien Béland [a, ✉], Shahab Jolani [b], François Pichette [c] & Jean-Sébastien Renaud [d]

[a]Faculté des sciences de l'éducation, Université de Montréal
[b]Department of Methodology and Statistics, CAPHRI, Maastricht University
[c]Département Sciences humaines, Lettre et Communications, Université TÉLUQ
[d]Faculté de médecine, Université Laval

Abstract ■ In Classical test theory, difficulty (p) and discrimination (d) are two item coefficients that are widely used to analyze and validate items in educational testing. However, test items are usually affected by missing data (MD), and little is known about the effect of methods for handling MD on these two coefficients. The current study compares several simple substitution (imputation) strategies for dichotomous items to better understand their impact on item difficulty and discrimination. We conducted a simulation study, followed by the analysis of a real data set of test items from a language test. Based on the root mean square errors (RMSE), person mean (PM) is the best overall replacement method for difficulty p and discrimination d. However, the analysis of bias coefficients and the analysis of real data show many similarities between most of the methods investigated to compute p while multiple imputation (MI) and complete cases (CC) seem to be the least biased methods to compute d.

Keywords ■ Classical test theory, item difficulty, item discrimination, missing data at random, educational testing. **Tools** ■ R.

✉ sebastien.beland@umontreal.ca

SB: na; *SJ*: 0000-0002-8508-0702; *FP*: na; *JSR*: 0000-0002-2816-0773

10.20982/tqmp.14.3.p180

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

Reviewers
■ One anonymous reviewer na

Introduction

Classical test theory (CTT) is a set of concepts and methods developed by various authors (e.g., Guilford, 1936; Gulliksen, 1950; Lord & Novick, 1968; Magnusson, 1967) based on the work of Spearman (Spearman, 1904, 1907, 1913). This theory is useful for item analysis, in order to collect information about item difficulty and item discrimination. Even if more modern measurement theories like Item Response Theory (IRT) are now extensively used, CTT remains popular because it presents important advantages. The most obvious one is the simplicity of CTT over IRT. For example, student ability is estimated by adding the number of correct answers to every item, in contrast with IRT, for which complex likelihood-based estimators must be used (Baker & Kim, 2004). Furthermore, CTT is implemented in most major statistical software (e.g., SPSS), and because it is the most taught and best known measurement theory, it is widely used, which facilitates comparisons of new results with those from previous studies. Another argument is the fact that CTT is less restrictive than IRT on the question of sample size and model assumptions.

The literature generally reports limitations for CTT (e.g., sample related coefficients) to justify the use of more recent psychometric approaches like IRT. However, DeVellis (2006) made this important statement:

> Some limitations of CTT are better documented in theory than in fact. I have personally observed instances in which scores on different versions of instruments, one based on CTT and the other based on more recent measurement models, have correlated with one another above 0.85. A correlation of that magnitude supports the conclusion that the 2 versions are fairly comparable. (p. S57)

In addition, Hays, Brown, Brown, Sprintzer, and Crall (2006) wrote that "CTT statistics are familiar and well understood and can help the user get oriented to the survey items and scales prior to estimating a more complex IRT model. We recommend further use of both approaches in tandem to evaluate survey items in future studies." (p. S67).

The use of CTT in educational testing

At least three reasons can explain the popularity of CTT in educational testing. First, this theory is reasonably easy to understand and, as mentioned earlier, it is generally available through popular software. Second, there exists an extensive didactic literature to help users analyze tests with CTT. Third, classroom assessment contexts deal with small sample sizes ($N < 80$), which are generally insufficient for modern approaches like IRT. And as mentioned by De Champlain (2010), this theory needs less restrictive assumptions. In this context, CTT is an interesting alternative to analyze the validity of educational tests.

The problem of missing data in educational testing

Students are expected to show their real ability in the context of classroom assessment, i.e. when doing educational tests. However, it is quite common to find missing data (MD) in educational testing. For example, a distracted student might intentionally skip some items, planning to answer them later on but forgetting to do so. A student can also deliberately avoid some items that bear sensitive content[1] or fail to answer difficult items.

What are the options when a student did not answer some test items? The choices available to test graders are displayed in the following decision tree (Figure 1).

In the presence of MD, is it possible to ask the student to answer all missing items? If so, the grader can directly assess the student's knowledge by questioning him/her, hence avoiding identification errors for the nature of the MD. If a student cannot be questioned again -often out of fairness for other students- the next step is to investigate the nature of the MD. If the grader knows precisely the nature of the MD (which is not always possible), he/she can justify the decision to ignore the unanswered items or to assign them a particular score. Substitution by zero, for example, is a common method used from the beginning of grade school, since teachers assume that it is for lack of knowing the correct answer that students do not respond. However, assuming lack of knowledge would be less justified, for example, when a high-achieving student simply skipped an item apparently due to distraction, while answering all the more difficult items correctly. In fact, such a student could perhaps provide the correct answer to the unanswered item spontaneously upon post-hoc questioning. Another possible situation of identifiable MD is when a student did not answer a series of questions simply for not knowing they were located on the back of a page. In short, when we know the nature of MD, principled methods like multiple imputation and direct likelihood methods have been shown to be suitable and can be used with confidence. Note that other simple methods are appropriate for particular situations; for example, Complete-case (CC) analysis tends to be appropriate in MCAR situations.

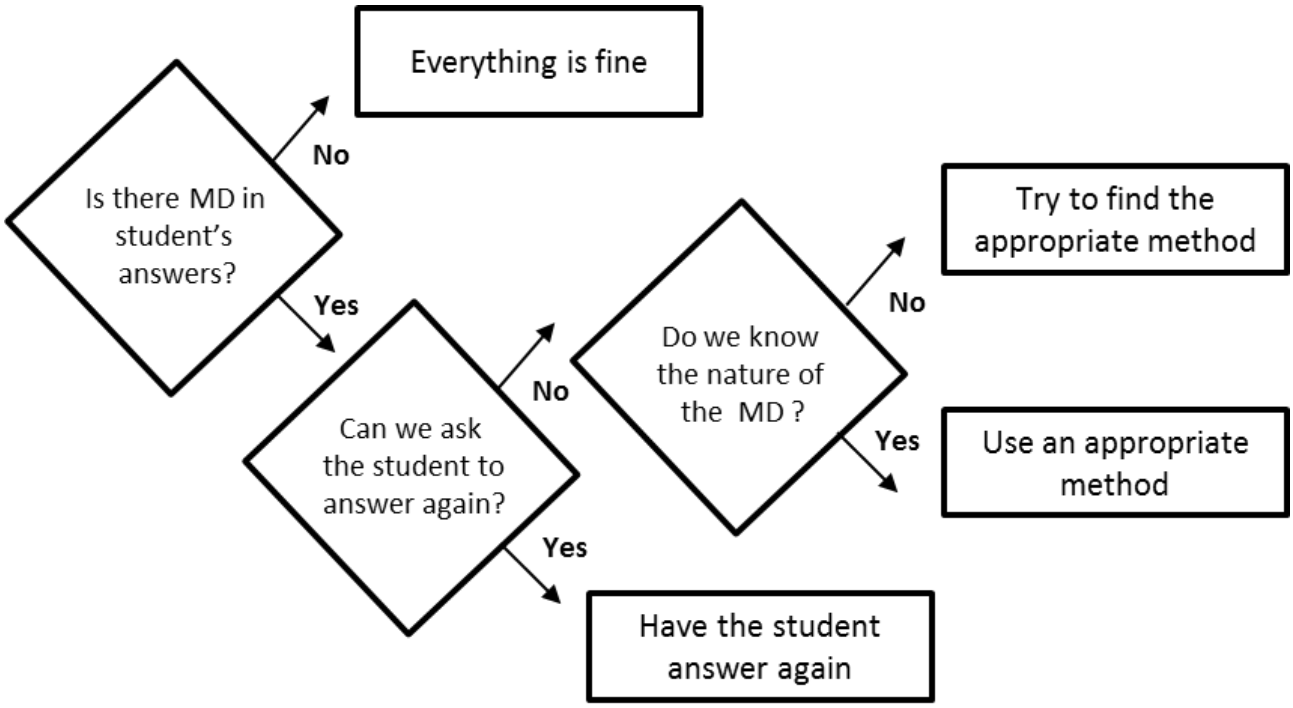
On the contrary, if the type of MD cannot be readily identified, there is no single universal method that can be assumed to work whichever MD type is in presence. It then becomes necessary to perform sensitivity analysis. Without knowing the nature of the MD in presence, the use of substitution techniques is important to minimize the occurrence of "biased estimates of parameters, loss of information, decreased statistical power, increased standard errors, and weakened generalizability of findings." (Dong & Peng, 2013, p. 1). These decisions can affect test validity, which is the most important quality of a test (Downing & Haladyna, 2009). As stated by Kane (2001), "validity is a property of the interpretations assigned to test scores, and these interpretations are considered valid if they are supported by convincing evidence" (p. 56). For example, the validity of a test gets distorted when MD underestimate the parameters in a psychometric model such as structural equation modeling. The presence of MD thus negatively impacts the interpretation of test scores and the psychometric properties of a test.

The impact of missing data in CTT: More research needed

Many authors have shown that when a data matrix contains MD, the statistical or measurement model in use can be biased (Allison, 2001; Bradlow & Thomas, 1998; Mackelprang, 1970; Rose, von Davier, & Xu, 2010; Schafer & Graham, 2002). In recent years, a great deal of attention has been devoted to the effect of MD on latent variable models; such was the case in exploratory and confirmatory factor analysis, structural equation modeling or IRT. For example, Finch (2008) investigated the effectiveness of several replacement methods on IRT difficulty and discrimination. Unfortunately, he was not able to highlight a superior method. For their part, Kamakura and Wedel (2000) developed a method to deal with MD in exploratory factor analysis, as well as Song and Lee (2002) in structural equation modeling. Other authors like Banks (2015) and Finch (2011) have discussed the case of MD on differential item functioning, and B. Zhang and Walker (2008) studied MD with person-fit statistics.

[1] See van Buuren (2012, p. 29) for a discussion about intentional and unintentional causes of MD.

Figure 1 ■ Grader's decision tree in the presence of missing data.



Surprisingly, very few studies explored the impact of MD on CTT item analysis. As an example, Béland, Pichette, and Jolani (2016), Enders (2004), and Sijtsma and Van der Ark (2003) studied the impact of many replacement methods on Cronbach's alpha coefficients. To the best of our knowledge, no studies have been conducted on the impact of substitution (imputation) methods on the estimates of difficulty and discrimination, which are used to assess the quality of items in the context of CTT (Livingston, 2012).

In CTT, item difficulty p is the proportion of correct answers per item. The value of this coefficient falls within a range from 0 to 1, and an item is considered easy when the p is high, and vice versa. Items with a p of .50 can be considered to be of moderate difficulty, while those at .85 and above can be considered easy, and those at .25 or less can be considered difficult (Hogan, Parent, & Stephenson, 2012; Laveault & Grégoire, 2014).

In addition, item discrimination d is the item's total biserial correlation (LeBlanc & Cox, 2017). Conceptually, d represents the difference between high performers and low performers and can be obtained with this formula:

$$d_i = \sqrt{p_i \times q_i} \times \frac{\mu_{1i} - \mu_{0i}}{\sigma_i}. \quad (1)$$

where μ_{1i} and μ_{0i} are the average scores of the students who answered correctly and incorrectly item i, respectively, σ_i is the standard deviation of item i, and pi and qi are respectively the proportion of students who answered the item correctly and incorrectly. The value of this d coefficient falls between -1 and 1. A low d-value indicates that a student who gets the item correct tends to perform poorly overall on the test, and vice versa. A value below 0 generally reflects a non-discriminating item, values between .01 and .09 a very poorly discriminating item, values between .10 and .19 a poorly discriminating item, values between .20 and .29 a moderately discriminating item, values between .30 and .39 an item showing good discrimination, and values over .39 an item with very good discrimination (Laveault & Grégoire, 2014; Nunnally & Bernstein, 1994; Schmeiser & Welch, 2006).

The treatment of missing data

Missingness mechanisms

The mechanisms that lead to missing data are usually classified into three categories (Rubin, 1976). The first category is missing completely at random (MCAR) and refers to situations where the absence of some data is entirely due to chance. The second category is missing at random (MAR), which occurs when missingness is entirely explained by the observed data rather than stemming from the miss-

ing data themselves. For concrete examples of educational testing situations for each of those two mechanisms, see Béland et al. (2016). Missingness can also be caused by the nature of the item, for example when questions are so intrusive or difficult that students elect not to provide an answer. This situation is known as missing not at random (MNAR). In the present study, we focus only on the first two mechanisms (i.e., MCAR and MAR). This strategy was also adopted in recent research (e.g., Q. Zhang & Wang, 2016, 4). The MNAR mechanism will not be considered, mostly because it involves very complex issues. To sum up, we assume that the data sets analyzed by educational researchers were obtained using psychometrically well-designed instruments that maximally reduce the chance of MNAR.

Substitution methods

Many methods are available to deal with missing data. In this section, we will focus on three categories: 1) deletion methods, 2) simple substitution methods, and 3) advanced methods. The interested readers can find an extended overview of methods in Allison (2001).

Among deletion methods, complete-case (CC) deletion consists of removing the data for any participant who provided missing values. For example, if a participant declines to answer at least one item, that person is discarded from the analysis. Another deletion strategy is the available case (AC) method. This method consists of eliminating missing data on a case-by-case analysis: the analyst only discards the missing answers on a test, and all available answers are used for the analysis.

Regarding simple substitution methods, various simple techniques can be used for a cognitive test consisting of right and wrong answers. Many graders consider missing data as incorrect answers for which the student gets "0". The idea behind this strategy is to sanction where there is no evidence of comprehension for an item. It is also remotely possible to consider a missing answer as being correct and assign it a score of "1". Among such rare instances, we could include the case mentioned earlier of a very strong student who forgot to answer the back side of an answer sheet.

Another strategy consists of substituting a missing value using mean-based methods. For example, the item mean (IM) of the observed cases is imputed for every missing datum:

$$IM = \sum_{i \in obs(j)} X_{ij} / \#obs\,(j) \tag{2}$$

where X_{ij} is the score of students i (i = 1, ..., n) to item j (j=1,. . . , J), and obs(j) denotes an item for which an answer is available. Another strategy is to replace the missing data by the participant's mean (PM) on the whole test. Here,

$$PM = \sum_{j \in obs(i)} X_{ij} / \#obs\,(i)\,. \tag{3}$$

where $obs(i)$ denotes respondents i who answered a specific question.

Winer (1971) proposed an alternative —called Winer imputation (W)— that combines IM and PM for missing data substitution:

$$W = \frac{IM + PM}{2}. \tag{4}$$

Three decades later, Huisman (2000) proposed a corrected item mean substitution (CIM):

$$CIM = \left(\frac{\#obs\,(j) \times PM}{\sum_{j \in obs(i)} IM} \right) \times IM \tag{5}$$

where "CIM replaces missing values by the item mean which is corrected for the 'ability' of the respondent, i.e., the score on the observed items of the respondent compared with the mean score on these items" (pp. 334-335).

van Ginkel, van der Ark, Sijtsma, and Vermunt (2007) reported the two-way imputation (TW):

$$TW = PM + IM - TM \tag{6}$$

where TM is the total mean of the test:

$$TM = \sum_{i,j \in obs} \sum X_{ij} / \#obs(i) \tag{7}$$

Finally, several advanced methods were developed to deal with missing data issues. A popular and powerful method is the likelihood-based method (ML, Allison, 2001), which is based on the product of complete data likelihood and incomplete data likelihood. The overall likelihood can then be maximized with different computation strategies to estimate parameters of interest.

Another powerful advanced method is multiple imputation (MI, Rubin, 1987). Like Peugh and Enders (2004), van Buuren (2012, p. 17) shows that multiple imputation comprises three main steps. The first step consists of imputing missing data from an incomplete data set to produce several complete (imputed) data sets. All the imputed data sets are different from one another in order to represent the uncertainty regarding which value to impute. This imputation step leads to multiple completed data sets, usually between three and five, although it is possible to increase it to a larger number, in the range of 50 or more. The second step is to perform the desired statistical analysis on each imputed data set (e.g., obtaining p, and d). In the third step, the results are pooled in order to obtain a single summary statistics. Following Rubin (1987), the pooled

estimate of the difficulties p (over imputations) is simply the arithmetic average over M estimates of p:

$$\bar{p} = \frac{1}{M} \times \sum_{m=1}^{M} p_k. \tag{8}$$

Schafer (2003) mentioned that MI and ML could lead to similar results under some specific conditions (e.g., using the same data set, a large sample size and a large number of imputations). This finding is also reported by Ibrahim, Chen, Lipsitz, and Herring (2005), Peugh and Enders (2004, see Table 4 on p. 547), and Kadengye, Ceulemans, and Van den Noortgate (2014). Finally, Q. Zhang and Wang (2013) showed that MI and ML performed well in the case of moderation analysis.

Because the ML and MI approaches are asymptotically equivalent, we only consider MI in this article. Moreover, MI can deal with a wider range of situations, which is the case in education testing. More importantly, MI "separates the solution of the missing data problem from the solution of the complete data problem" (van Buuren, 2012, p. 18). As the readers will see, this can lead to a better understanding of many research questions.

Aim of this study

The aim of this study is to use a simulation study as well as a real data set to compare the precision of the estimate of p and d across ten common fill-in methods for handling missingness under MCAR and MAR mechanisms for normal-size dichotomous data sets in education testing. As we mentioned earlier, the MNAR mechanism involves highly complex issues where the reason for missing data must also be modeled, and is therefore not considered here.

Method

Ten methods are compared to investigate their impact on p and d coefficients: multiple imputation (MI), corrected item mean (CIM), Winer imputation (W), two-way imputation (TW), replacement by the person's mean (PM), replacement by the item's mean (IM), replacement by the total mean (TM), replacement by "0", replacement by "1", and only complete cases (CC).

Study 1: Simulation study

Our procedure was based on the collection and analysis of dichotomous data (e.g., correct/incorrect answers). We used the sim.rasch function from the 'psych 1.8.4' R package (Revelle, 2018) to create dichotomous data sets from Rasch models where the ability and difficulty parameters were generate from a $N(0, 1)$. The data sets contained either 100 or 500 participants with two test lengths (20 and 60). For this study, we chose two percentages of missing answers: 0.05 and 0.20. Missing values were then obtained under MCAR and MAR mechanisms. For MCAR, missing values were generated at random in every dichotomous data matrix. In the case of the MAR mechanism, we adopted the strategy suggested by van Buuren, Brand, Groothuis-Oudshoorn, and Rubin (2006). This procedure ensures that, for each participant, the probability for an item to be missing only depends on the observed items for that participant. Finally, each item presents a similar number of missing answers.

The bias and root mean square error (RMSE) will be reported for each combination (number of items × number of participants × percentage of missing answers). To obtain these statistics, we first started by generating full dichotomous matrices (i.e., without missing values) and the corresponding p and d values were obtained. We considered these estimated values to be the true values for p and d. Next, missing values were inserted in these matrices and the p and d quantities were estimated after applying a missing data treatment method (e.g., item mean imputation). To compute bias, the mean difference was calculated between the true values of p and d and their estimated values (after applying a missing data treatment method) across 1000 replications. For RMSE, the following formula was used:

$$RMSE = \sqrt{bias^2 + SD^2} \tag{9}$$

where SD is the standard deviation of p and d per method that are obtained from 1000 replications.

Study 2: Real data analysis

We used the TCALS-II (Test de Classement en Anglais, Langue Seconde - Version II [Placement Test of English as a Second Language]) to test the impact of the replacement methods. This test assesses the English competence of French-speaking students entering college. The TCALS-II contains 85 multiple-choice items divided into eight subgroups. The data matrix under consideration are the complete responses of students to these 85 items over three different times: $N = 1372$ (1998), $N = 1279$ (2004), and $N = 1835$ (2008) at a College located in Western Quebec. These years were selected because of their unidimensionality and mutually exclusive items.

We generated 0.20 of missing answers under MCAR and MAR mechanisms where each item presents almost the same number of missing answers. We then substituted every missing datum using the ten above-mentioned methods. The results from each substitution method were compared with the full data matrix before introducing missing data. Like Study 1, that original full matrix is used as reference. Finally, it is important to mention that we choose the TCALS-II because it respects the assumptions of unidimensionality and relative independence between items (Raîche, 2002). This is a crucial point that ensures us to

make credible comparisons with Study 1 because we make use of the Rasch model to generate data matrices.

Results

Study 1: Simulation study

The next output displays the results for p. First, Table 1 reports simulations with 20 items. Except for "0" and "1", the RMSE values are smaller when the matrix size increases ($N = 500$). The low RMSE values for PM and TM suggest that they are the best overall methods. However, it is important to notice that RMSE values can also be low for IM, W, TW, CIM and MI.

The bias coefficient values are generally low, which indicates consistency in item difficulty between the replacement methods that are compared. This observation holds true especially when the rate of missingness is equal to 0.05. However, we have to point out that the "0" and "1" replacement methods are clearly the most biased ones.

Table 2, based on 60-item simulations, shows the tendencies observed in the previous table. This time again, a larger N leads to a smaller RMSE, and the methods that yield the lowest RMSE are respectively PM and TM. Finally, the bias is systematically higher for replacement by "0" and "1".

Tables 3 and 4 present the results for d, based on 20 and 60 items respectively. To begin, it is important to highlight that the RMSE and bias values we obtained are higher for d than they were for p (as shown previously in Table 1 and 2). Furthermore, RMSE values from those two tables become smaller as the number of items increases.

As we can see in Table 3, RMSE are at their smallest value when PM and W are in use. The study of bias shows that MI, TW and CC are the least biased methods when $N = 100$, while MI and CC present the smallest bias when $N = 500$. Finally, we see that "0" and "1" are not adequate replacement methods for r.

The result for 60 items can be synthesized as follows. First of all, the RMSE are generally lower when the number of item rises. Second, PM and TW are respectively the replacement methods with the lowest RMSE. Finally, bias coefficients suggest that CC, CIM, and MI are respectively the best replacement methods.

Study 2: Real data analysis

Table 5 shows the results under an MCAR mechanism with 0.20 of MD. When compared to the reference, the p values were similar for MI, CIM, TW, W, PM, IM, TM and CC. This is in accordance with some results of Study 1, where the difference between the bias coefficients of these methods can be very small (e.g., MCAR & rate=0.05 and MCAR & rate=0.20 when $N = 500$). Furthermore, our previous results also show that the "0" and "1" replacement methods are not precise. For the d values, MI and CC are the most interesting methods, which is in accordance with many bias coefficient results of Table 3 and Table 4.

Again, Table 6 shows that the difficulties are relatively similar for MI, CIM, TW, W, PM, IM and TM. Contrary to Table 1, CC replacement method is now less powerful to recover the references. Finally, MI and CC are the best options for item discrimination.

Synthesis

Over the 1,000 generated matrices, the RMSE are slightly higher when $N = 100$ and for item discrimination d. Based on our overall results, PM is a slightly better replacement method for difficulty p and for discrimination d. However, bias coefficients from Study 1 and the results from Study 2 show many similarities between CC, MI, CIM, TW, W, PM, IM, and TM for difficulty p, while MI and CC appear to be the most appropriate methods for d. Finally, it may be a bad decision to use "0" and "1" replacement methods, given that their RMSE and bias coefficients are the highest ones.

Discussion

In our study, the MI substitution method proves very efficient based on the bias coefficients of Table 1 to 4. In other contexts, Béland et al. (2016), Schafer and Graham (2002) and van Buuren (2012), among others, also showed that MI is among the better approaches for dealing with missing data.

Mean-based substitution methods do not enjoy a good reputation. According to Enders (2010, p. 43): "simulation studies suggest that mean imputation is possibly the worst missing data handling method available. Consequently, in no situation is mean imputation defensible, and you should absolutely avoid this approach". van Buuren (2012) also mentions that:

> mean imputation is a fast and simple fix for the missing data. However, it will underestimate the variance, disturb the relations between variables, bias almost any estimate other than the mean and bias the estimate of the mean when data are not MCAR. Mean imputation should perhaps only be used as a rapid fix when a handful of values are missing, and it should be avoided in general. (p.11)

An important observation that can be made from Table 1 to Table 6 is that mean-based procedures are not always as useless as suggested by many authors. For example, in Study 2, CIM, TW, W, PM, IM and TM are all very close to the reference for p. Furthermore, Béland et al. (2016) also found that IM, TM, and W can minimize the impact of miss-

Table 1 ■ RMSE and bias for difficulty p (20 items)

	CC	0	1	TM	IM	PM	W	TW	CIM	MI
					$N = 100$					
					MCAR & rate=0.05					
RMSE	0.0529	0.0555	0.0555	0.0491	0.0515	0.0489	0.0502	0.0513	0.0514	0.0513
Bias	0.0003	-0.0249	0.0251	0.0001	0.0001	-0.0001	0.0000	-0.0001	-0.0001	0.0001
					MAR rate=0.05					
RMSE	0.0527	0.0563	0.0567	0.0488	0.0512	0.0487	0.0499	0.0511	0.0511	0.0509
Bias	0.0015	-0.0249	0.0257	0.0004	0.0004	-0.0005	0.0000	-0.0005	-0.0005	0.0002
					MCAR rate=0.20					
RMSE	0.0651	0.1107	0.111	0.0463	0.0566	0.0459	0.051	0.0561	0.0566	0.0553
Bias	0.0001	-0.0998	0.1002	0.0002	0.0003	0.0004	0.0003	0.0004	0.0003	0.0003
					MAR rate=0.20					
RMSE	0.0664	0.1136	0.1164	0.0458	0.0563	0.0456	0.0506	0.056	0.056	0.0551
Bias	0.0104	-0.1015	0.1042	0.0015	0.0016	-0.0036	-0.0010	-0.0036	-0.0036	0.0001
					$N = 500$					
					MCAR rate=0.05					
RMSE	0.0235	0.0333	0.0334	0.0219	0.023	0.0219	0.0224	0.0229	0.023	0.023
Bias	0.0000	-0.0250	0.0250	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0001
					MAR rate=0.05					
RMSE	0.0236	0.0336	0.0342	0.0218	0.0229	0.0218	0.0223	0.0228	0.0229	0.0228
Bias	0.0013	-0.0250	0.0256	0.0003	0.0003	-0.0006	-0.0001	-0.0006	-0.0006	0.0001
					MCAR rate=0.20					
RMSE	0.0287	0.1022	0.1022	0.0206	0.0251	0.0204	0.0227	0.0249	0.0252	0.025
Bias	-0.0001	-0.1000	0.1000	0.0001	0.0001	0.0002	0.0001	0.0002	0.0002	0.0001
					MAR rate=0.20					
RMSE	0.031	0.1037	0.1068	0.0205	0.0251	0.0205	0.0226	0.025	0.0251	0.025
Bias	0.0106	-0.1011	0.1043	0.0020	0.0020	-0.0031	-0.0006	-0.0031	-0.0031	0.0002

ing data on Cronbach's alpha when analyzing small sample sizes. Finally, Sijtsma and Van der Ark (2003) show that simple replacement methods like PM and TW display small bias when analyzing incomplete data matrices with Cronbach's alpha.

Limitations

Obviously, the current study is not without limitations. First, we only analyzed the case of unidimensional and mutually exclusive items in our simulation study. Although this setting is of interest in educational testing, there are many situations in which scientists deal with multidimensional data matrices. Second, the simulation design can have an impact on the performance of some replacement method. Third, the item range under investigation in this article is limited to 60 items. This choice is pertinent for an exploratory study such as our application, but tests can be longer in real-life situations. Finally, we excluded the possibility of MNAR mechanisms, which suggests that the current study only informs us about the question of "random missing data mechanisms".

Conclusion

In test situations, graders generally assign a score of "0" to a student who failed to answer an item. Our results suggest that this strategy is quite misguided when analyzing the psychometric qualities of a test, even when the rate of MD is very low (i.e., 0.05). In Study 1, the RMSE coefficients suggest that PM is the best overall method for computing p as well as for d. However, the substitution methods CC, MI, CIM, TW, W, PM, IM, and TM generally lead to similar bias results for p. In the case of d, MI and CC present the smallest bias.

More studies are needed to understand the impact of missing data on item analysis. Here are five suggested avenues for further research. First, our study stresses the need to investigate how these results can be extended to multidimensional data matrices or to dependency between dichotomous items. Second, our study raises interest in investigating the effect of MD using polytomous items, such as data obtained from Likert scales. Third, the effect of MD on the estimation of true ability under CTT can be a stimulating avenue for future studies. Fourth, this study consid-

Table 2 ■ RMSE and bias for difficulty p (60 items)

	CC	0	1	TM	IM	PM	W	TW	CIM	MI
					$N = 100$					
					MCAR & rate=0.05					
RMSE	0.0527	0.0553	0.0553	0.0488	0.0512	0.0487	0.0500	0.0511	0.0512	0.0500
Bias	0.0000	-0.0250	0.0250	0.0000	0.0000	-0.0001	0.0000	-0.0001	-0.0001	0.0000
					MAR & rate=0.05					
RMSE	0.0529	0.0563	0.0567	0.0490	0.0514	0.0489	0.0501	0.0513	0.0513	0.0502
Bias	0.0006	-0.0248	0.0252	0.0002	0.0002	-0.0002	0.0000	-0.0002	-0.0002	0.0001
					MCAR & rate=0.20					
RMSE	0.0646	0.1106	0.1107	0.0456	0.0560	0.0452	0.0505	0.0556	0.0558	0.0495
Bias	0.0000	0.0000	-0.0999	0.1001	0.0000	0.0000	0.0000	0.0000	-0.0001	-0.0001
					MAR & rate=0.20					
RMSE	0.0655	0.1118	0.1155	0.0457	0.0563	0.0453	0.0506	0.0559	0.0557	0.0494
Bias	0.0000	0.0085	-0.0993	0.1034	0.0026	-0.0009	0.0008	-0.0009	-0.0010	0.0016
					$N = 500$					
					MCAR & rate=0.05					
RMSE	0.0236	0.0333	0.0334	0.0219	0.0230	0.0218	0.0224	0.0229	0.0229	0.0228
Bias	0.0001	-0.0250	0.0250	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
					MAR & rate=0.05					
RMSE	0.0236	0.0335	0.0338	0.0219	0.0230	0.0219	0.0224	0.0230	0.0230	0.0229
Bias	0.0007	-0.0247	0.0251	0.0002	0.0002	-0.0002	0.0000	-0.0002	-0.0002	0.0000
					MCAR & rate=0.20					
RMSE	0.0288	0.1022	0.1022	0.0203	0.0249	0.0201	0.0224	0.0247	0.0248	0.0244
Bias	-0.0001	-0.1000	0.1000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
					MAR & rate=0.20					
RMSE	0.0302	0.1018	0.1056	0.0204	0.0251	0.0202	0.0225	0.0249	0.0248	0.0245
Bias	0.0084	-0.0992	0.1030	0.0024	0.0024	-0.0012	0.0006	-0.0011	-0.0012	0.0002

ered MCAR and MAR missingness mechanisms. However, there are situations in educational testing where these assumptions are (clearly) violated. Future research could investigate the effect of MNAR mechanisms on item difficulty and discrimination in the context of educational testing. Fifth, we are eager to better understand why the simple mean-based substitution methods work well in the context of unidimensional dichotomous items.

References

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

Baker, F. B., & Kim, S.-h. (2004). *Item response theory: Parameter estimation techniques (2nd ed.)* New Jersey: Dekker.

Banks, K. (2015). An introduction to missing data in the context of differential item functioning. *Practical Assessment, Research & Evaluation*, *20*. Retrieved from http://pareonline.net/getvn.asp?v=20&n=12

Béland, S., Pichette, F., & Jolani, S. (2016). Impact of simple treatment of missing data on cronbach's a coefficient. *The Quantitative Methods for Psychology*, *12*, 57–73. doi:10.20982/tqmp.12.1.p057

Bradlow, E. T., & Thomas, N. (1998). Item response theory models applied to data allowing examinee choice. *Journal of Educational and Behavioral Statistics*, *23*, 236–243.

De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education*, *44*, 109–117.

DeVellis, R. F. (2006). Classical test theory. *Medical Care Research and Review*, *44*, S50–S59.

Dong, Y., & Peng, C. Y. J. (2013). *Principled missing data methods for researchers*. Plus, 2 , Available online: Springer. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3701793/pdf/40064_2013_Article_296.pdf

Downing, S. M., & Haladyna, T. M. (2009). Validity and its threats. In S. M. Downing & R. Yudkowsky (Eds.), *Assessment in health professions education* (pp. 21–55). New York: Routledge.

Enders, C. K. (2004). The impact of missing data on sample reliability estimates: Implications for reliability reporting practices. *Educational and Psycho-*

Table 3 ■ RMSE and bias for discrimination d (20 items)

	CC	0	1	TM	IM	PM	W	TW	CIM	MI
					$N = 100$					
					MCAR & rate=0.05					
RMSE	0.1305	0.1313	0.1320	0.1246	0.1246	0.1185	0.1213	0.1189	0.1192	0.1275
Bias	0.0006	0.0537	0.0517	-0.0052	-0.0051	0.0009	-0.0014	0.0019	0.0034	-0.0015
					MAR & rate=0.05					
RMSE	0.1302	0.1350	0.1310	0.1233	0.1233	0.1160	0.1193	0.1164	0.1169	0.1271
Bias	-0.0055	0.0561	0.0453	-0.0089	-0.0088	-0.0005	-0.0036	0.0007	0.0025	-0.0025
					MCAR & rate=0.20					
RMSE	0.1598	0.1634	0.1631	0.1270	0.1270	0.1151	0.1169	0.1175	0.1181	0.1401
Bias	0.0013	0.1144	0.1154	-0.0275	-0.0273	0.0410	0.0249	0.0441	0.0453	-0.0109
					MAR & rate=0.20					
RMSE	0.1655	0.1820	0.1513	0.1290	0.1291	0.1179	0.1176	0.1204	0.1240	0.1439
Bias	-0.0144	0.1379	0.0861	-0.0341	-0.0339	0.0458	0.0262	0.0493	0.0555	-0.0114
					$N = 500$					
					MCAR & rate=0.05					
RMSE	0.0570	0.0734	0.0741	0.0547	0.0547	0.0520	0.0532	0.0522	0.0534	0.0561
Bias	0.0005	0.0512	0.0520	-0.0055	-0.0054	0.0016	-0.0005	0.0024	0.0092	-0.0003
					MAR & rate=0.05					
RMSE	0.0575	0.0775	0.0708	0.0550	0.0550	0.0516	0.0529	0.0520	0.0555	0.0566
Bias	-0.0054	0.0556	0.0449	-0.0087	-0.0087	0.0006	-0.0025	0.0017	0.0154	-0.0006
					MCAR & rate=0.20					
RMSE	0.0711	0.1258	0.1265	0.0617	0.0617	0.0653	0.0579	0.0660	0.1533	0.0645
Bias	0.0008	0.1146	0.1154	-0.0264	-0.0263	0.0432	0.0258	0.0441	0.1456	-0.0016
					MAR & rate=0.20					
RMSE	0.0735	0.1472	0.1022	0.0646	0.0645	0.0666	0.0569	0.0673	0.1579	0.0652
Bias	-0.0158	0.1375	0.0864	-0.0338	-0.0337	0.0456	0.0251	0.0466	0.1503	-0.0028

logical Measurement, *64*, 419–436. doi:10 . 1177 / 0013164403261050

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford.

Finch, H. (2008). Estimation of item response theory parameters values in the presence of missing data. *Journal of Educational Measurement*, *45*, 225–246.

Finch, H. (2011). The use of multiple imputation for missing data in uniform dif analysis: Power and type I error rates. *Applied Measurement in Education*, *24*, 281–301. doi:http://dx.doi.org/10.1080/08957347.2011.607054

Guilford, J. P. (1936). *Psychometric methods*. New York: McGraw-Hill.

Gulliksen, H. (1950). *Theory of mental tests*. Hoboken, NJ: John Wiley & Sons Inc. Retrieved from http://content.apa.org/books/2009-12806-000

Hays, R. D., Brown, J., Brown, L. U., Sprintzer, K. L., & Crall, J. J. (2006). Classical test theory and item response theory analyses of multi-item scales assessing parents perception of their children's dental care. *Medical Care*, *44*, S60–S68.

Hogan, T. P., Parent, N., & Stephenson, R. (2012). *Introduction à la psychométrie*. Montréal: Chenelière.

Huisman, M. (2000). Imputation of missing item responses: Some simple techniques. *Quality & Quantity*, *34*, 331–351. doi:10.1023/A:100478223006

Ibrahim, J. G., Chen, M.-h., Lipsitz, S. R., & Herring, A. H. (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, *100*, 332–346.

Kadengye, D. T., Ceulemans, E., & Van den Noortgate, W. (2014). Direct likelihood analysis and multiple imputation for missing item scores in multilevel cross-classification educational data. *Applied Psychological Measurement*, *38*, 61–80. doi:10 . 1177 / 0146621613491138

Kamakura, W. A., & Wedel, M. (2000). Factor analysis and missing data. *Journal of Marketing Research*, *37*, 490–498. doi:10.1111/j.1749-6632.1965.tb11694.x

Kane, M. T. (2001). So much remains the same: Conception and status of validation in standard setting methods. In G. J. Cizek (Ed.), *Setting performance standards:*

Table 4 ■ RMSE and bias for discrimination d (60 items)

	CC	0	1	TM	IM	PM	W	TW	CIM	MI
					$N = 100$					
					MCAR & rate=0.05					
RMSE	0.1270	0.1260	0.1262	0.1219	0.1219	0.1167	0.1192	0.1168	0.1168	0.1214
Bias	0.0003	0.0485	0.0484	-0.0060	-0.0060	-0.0099	-0.0060	-0.0096	-0.0088	-0.0078
					MAR & rate=0.05					
RMSE	0.1265	0.1291	0.1273	0.1207	0.1207	0.1146	0.1174	0.1147	0.1148	0.1208
Bias	-0.0057	0.0479	0.0439	-0.0107	-0.0106	-0.0144	-0.0099	-0.0140	-0.0133	-0.0107
					MCAR & rate=0.20					
RMSE	0.1563	0.1526	0.1531	0.1283	0.1284	0.1102	0.1173	0.1104	0.1084	0.1276
Bias	0.0005	0.1048	0.1051	-0.0300	-0.0298	0.0127	0.0158	0.0136	0.0024	-0.0420
					MAR & rate=0.20					
RMSE	0.1584	0.1631	0.1458	0.1288	0.1288	0.1067	0.1143	0.1071	0.1060	0.1291
Bias	-0.0155	0.1170	0.0857	-0.0398	-0.0396	0.0092	0.0118	0.0104	0.0013	-0.0487
					$N = 500$					
					MCAR & rate=0.05					
RMSE	0.0560	0.0706	0.0700	0.0541	0.0541	0.0516	0.0526	0.0516	0.0513	0.0546
Bias	0.0002	0.0484	0.0478	-0.0058	-0.0058	-0.0034	-0.0015	-0.0034	-0.0020	-0.0014
					MAR & rate=0.05					
RMSE	0.0560	0.0717	0.0690	0.0541	0.0541	0.0510	0.0519	0.0510	0.0508	0.0545
Bias	-0.0049	0.0483	0.0439	-0.0097	-0.0097	-0.0054	-0.0038	-0.0053	-0.0003	-0.0017
					MCAR & rate=0.20					
RMSE	0.0683	0.1156	0.1160	0.0623	0.0623	0.0506	0.0535	0.0507	0.1150	0.0602
Bias	0.0000	0.1047	0.1051	-0.0294	-0.0293	-0.0177	0.0174	0.0179	0.1048	-0.0066
					MAR & rate=0.20					
RMSE	0.0708	0.1281	0.1008	0.0665	0.0665	0.0491	0.0521	0.0493	0.1171	0.0611
Bias	-0.0148	0.1176	0.0863	-0.0381	-0.0381	0.0161	0.0152	0.0163	0.1070	-0.0069

Concepts, methods, and perspectives (pp. 53–88). Mahwah, NJ: Erlbaum.

Laveault, D., & Grégoire, J. (2014). *Introduction aux théories des tests en psychologie et en sciences de l'éducation (3ième edition)*. Bruxelle: De Boeck.

LeBlanc, V., & Cox, M. A. A. (2017). Interpretation of the point-biserial correlation coefficient in the context of a school examination. *The Quantitative Methods for Psychology*, *13*, 46–56. doi:10.20982/tqmp.13.1.p046

Livingston, S. A. (2012). Item analysis. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 421–441). New York: Routledge.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley.

Mackelprang, A. J. (1970). Missing data in factor analysis and multiple regression. *Midwest Journal of Political Science*, *14*, 493–505.

Magnusson, D. (1967). *Test theory*. Reading: Addison-Wesley.

Nunnally, J., & Bernstein, L. (1994). *Psychometric theory*. New York: McGraw-Hill.

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*, 525–556. doi:10 . 3102 / 00346543074004525

Raîche, G. (2002). *Le dépistage du sous-classement aux tests de classement en anglais, langue seconde, au collégial*. Gatineau, Québec: Collège de l'Outaouais.

Revelle, W. (2018). Psych: Procedures for personality and psychological research (Version 1.8.4). Retrieved from https://CRAN.R-project.org/package=psych

Rose, N., von Davier, M., & Xu, X. (2010). *Modeling non-ignorable missing data with item response theory (irt)*. Princeton: Educational Testing Service.

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

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Schafer, J. L. (2003). Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica neerlandica*, *57*, 19–35.

Table 5 ■ Item difficulty and item discrimination for the TCALS-II under MCAR

	1998		2004		2008	
	p	d	p	d	p	d
Reference	0.7763	0.5837	0.7720	0.5811	0.7576	0.5647
MI	0.7746	0.5807	0.7711	0.5805	0.7561	0.5596
CIM	0.7747	0.6898	0.7712	0.6867	0.7565	0.6622
TW	0.7748	0.701	0.7712	0.6982	0.7566	0.6617
W	0.7747	0.6547	0.7713	0.6506	0.7559	0.6166
PM	0.7748	0.6779	0.7710	0.6727	0.7568	0.6381
IM	0.7747	0.4684	0.7716	0.4657	0.755	0.4551
TM	0.7747	0.4677	0.7715	0.4652	0.7551	0.4538
1	0.8192	0.6341	0.8167	0.6329	0.8035	0.6147
0	0.6215	0.5489	0.6189	0.5474	0.6059	0.5334
CC	0.7745	0.5898	0.7722	0.5795	0.7523	0.5699

Table 6 ■ Item difficulty and item discrimination for the TCALS-II under MAR

	1998		2004		2008	
	p	d	p	d	p	d
Reference	0.7763	0.5837	0.7720	0.5811	0.7576	0.5647
MI	0.7754	0.5857	0.7718	0.5796	0.7564	0.5618
CIM	0.7759	0.7085	0.7722	0.7066	0.7565	0.6847
TW	0.7761	0.7082	0.7722	0.6932	0.7564	0.6616
W	0.7790	0.6663	0.7751	0.6464	0.7595	0.6175
PM	0.7764	0.6888	0.7723	0.6663	0.7561	0.6362
IM	0.7815	0.4754	0.7779	0.4775	0.7629	0.4705
TM	0.7819	0.4739	0.778	0.4764	0.7626	0.4701
1	0.8266	0.648	0.8202	0.6444	0.8096	0.634
0	0.6217	0.5312	0.6301	0.5248	0.6117	0.5112
CC	0.7914	0.5783	0.7875	0.5805	0.7736	0.5589

Schafer, J. L., & Graham, J. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177. doi:10.1037//1082-989X.7.2.147

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In *Educational measurement* (pp. 307–353). Westport: Praeger.

Sijtsma, K., & Van der Ark, L. A. (2003). Investigation and treatment of missing item scores in test and questionnaire data. *Multivariate Behavioral Research*, *38*, 505–528. doi:10.1207/s15327906mbr3804_4

Song, X.-y., & Lee, S. Y. (2002). Analysis of structural equation model with ignorable missing continuous and polytomous data. *Psychometrika*, *67*, 261–288. doi:10.1007/BF02294846

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*, 72–101.

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *The American Journal of Psychology*, *18*, 161–169.

Spearman, C. (1913). Correlations of sums or differences. *British Journal of Psychology*, *5*, 417–426.

van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton: CRC.

van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, K., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, *76*, 1049–1064. doi:10.1080/10629360600810434

van Ginkel, J. R., van der Ark, L. A., Sijtsma, K., & Vermunt, J. K. (2007). Two-way imputation: A Bayesian method for estimating missing scores in tests and questionnaires, and an accurate approximation. *Computational Statistics and Data Analysis*, *51*, 4013–4027.

Winer, B. J. (1971). *Statistical principles in experimental design*. New York: McGraw-Hill.

Zhang, B., & Walker, C. M. (2008). Impact of missing data on person model fit and person trait estimation. *Applied Psychological Measurement*, *32*, 466–480. doi:10.1177/0146621607307692

Zhang, Q., & Wang, L. (2013). Methods for mediation analysis with missing data. *Psychometrika*, *78*, 154–184. doi:10.1007/s11336-012-9301-5

Zhang, Q., & Wang, L. (2016). Moderation analysis with missing data in the predictors. *22*, 649–666. doi:10.1037/met0000104

Appendix: R code for the analysis of a single data matrix containing missing values

```
ctt <- function(y){
  temp <- reliability(y)
  out <- rbind(difficulty = temp$itemMean, discrimination = temp$bis)
  return(out)
}

# Multiple imputation
multiple.impute <- function(data){
    imp <- mice(data, print = F)
    mires <- array(NA, dim = c(2, ncol(data), imp$m))
    for (i in 1:imp$m) mires[,,i] <- ctt(complete(imp, i))
    result <- apply(mires, 1:2, mean)
    return(result)
}

# single imputation methods
single.impute <- function(data){
  out <- array(NA, dim = c(2, ncol(data), 8),
               dimnames = list(c("dif", "dis"), NULL,
                               c("Zero", "One", "TM", "IM", "PM", "Winer", "Two-way
   ", "CIM")))
  temp2 <- temp1 <- temp <- data
    # zero replacement
    temp[is.na(data)] <- 0
    out[,,"Zero"] <- ctt(temp)
    # one replacement
    temp[is.na(data)] <- 1
    out[,,"One"] <- ctt(temp)
    # overall mean imputation
    temp[is.na(data)] <- mean(data, na.rm = T)
    out[,,"TM"] <- ctt(temp)
    # item's mean imputation
    i.mean <- colMeans(data, na.rm = T)
    for (j in 1:ncol(data)){ temp[,j][is.na(data[,j])] <- i.mean[j] }
    out[,,"IM"] <- ctt(temp)
    # participant's mean imputation
    p.mean <- rowMeans(data, na.rm = T)
    for (i in 1:nrow(data)){ temp[i,][is.na(data[i,])] <- p.mean[i] }
    out[,,"PM"] <- ctt(temp)
    # other sigle imputation methods
    for (i in 1:nrow(data)){
      for (j in 1:ncol(data)){
        temp[i,j][is.na(data[i,j])] <- (p.mean[i] + i.mean[j])/2
        temp1[i,j][is.na(data[i,j])] <- p.mean[i] + i.mean[j] - mean(data, na.rm =
   T)
        temp2[i,j][is.na(data[i,j])] <- ((sum(!is.na(data[i,]))*p.mean[i])/(sum(
   i.mean[!is.na(data[i,])])))*i.mean[j]
```

```
      }
    }
  # average of item and participant's mean (Winer's method)
  out[,,"Winer"] <- ctt(temp)
  # Two-way imputation method (Sijtsma's method); PM + IM - OM
  out[,,"Two-way"] <- ctt(temp1)
  # corrected item mean imputation
  out[,,"CIM"] <- ctt(temp2)
  return(out)
}

# main program
prog <- function(data){
  suppressWarnings(if(!require(mice)) paste("Install␣the␣'mice'␣package"))
  suppressWarnings(if(!require(CTT)) paste("Install␣the␣'CTT'␣package"))
  data <- as.matrix(data)
  # 'out' object contains the results
  # first layer corresponds to item difficulty and discrimination (2)
  # second layer corresponds to the number of items in the matrix
  # third layer corresponds to the number of methods (10 methods)
  out <- array(NA, dim = c(2, ncol(data), 10),
               dimnames = list(c("dif", "dis"), NULL,
                               c("CC", "Zero", "One", "TM", "IM", "PM", "Winer", "
    Two-way", "CIM", "MI")))
  # Methods
  # Complete case analysis
  out[,,1] <- ctt(data)
  # Single imputation
  out[,,2:9] <- single.impute(data)
  # Multiple imputation
  out[,,10] <- multiple.impute(data)
  return(out)
}
```

Citation

Béland, S., Jolani, S., Pichette, F., & Renaud, J.-S. (2018). Impact of simple substitution methods for missing data on classical test theory difficulty and discrimination. *The Quantitative Methods for Psychology*, *14*(3), 180–192. doi:10.20982/tqmp.14.3.p180

Received: 26/07/2017 ~ Accepted: 02/08/2018