Imputation Methods to Deal With Missing Responses in Computerized Adaptive Multistage Testing

Educational and Psychological

Measurement

2019, Vol. 79(3) 495–511

© The Author(s) 2018

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0013164418805532

journals.sagepub.com/home/epm

Dee Duygu Cetin-Berber¹, Halil Ibrahim Sari²
and Anne Corinne Huggins-Manley¹

Abstract

Routing examinees to modules based on their ability level is a very important aspect in computerized adaptive multistage testing. However, the presence of missing responses may complicate estimation of examinee ability, which may result in misrouting of individuals. Therefore, missing responses should be handled carefully. This study investigated multiple missing data methods in computerized adaptive multistage testing, including two imputation techniques, the use of full information maximum likelihood and the use of scoring missing data as incorrect. These methods were examined under the missing completely at random, missing at random, and missing not at random frameworks, as well as other testing conditions. Comparisons were made to baseline conditions where no missing data were present. The results showed that imputation and the full information maximum likelihood methods outperformed incorrect scoring methods in terms of average bias, average root mean square error, and correlation between estimated and true thetas.

Keywords

Missing data, imputation, computerized adaptive multistage testing

Introduction

Computerized adaptive multistage testing (ca-MST) is a form of test administration that makes use of stages, or divisions, of a test. In each stage there are preconstructed

¹University of Florida, Gainesville, FL, USA

²7 Aralik University, Kilis, Turkey

Corresponding Author:

Dee Duygu Cetin-Berber, School of Human Development and Organizational Studies in Education, College of Education, University of Florida, Norman Hall G514B, Gainesville, FL 32611, USA.

Email: dcetinberber@ufl.edu

groups of items called modules (e.g., a mini test consisting of 5 or 10 items), and the modules often vary in difficulty. In ca-MST, the test administration begins with a routing module, which is the initial module administered to an examinee. Based on the examinee's performance on the routing module, the computer selects another module for the next stage of test administration (Luecht, Brumfield, & Breithaupt, 2006). This module-level adaptation continues until an examinee completes all stages, with the number of total stages often predetermined by the test developers.

Tests administered through module-level adaptation may allow examinees to skip items in a module, or more generally to respond to module test items in the order of their choice. This is in contrast to tests administered through item-level adaptation, which most often require a response in order to proceed in the test. However, the ability to skip items in ca-MST may increase the likelihood of having test data that contain missing item responses. It is clear from the measurement literature that standard psychometric procedures for analyzing item response data may produce biased ability estimates in the presence of missingness (e.g., DeMars, 2002; Finch, 2008).

In ca-MST, one of the simplest approaches to deal with missing responses is treating all missing values as incorrect. For example, this is the approach used by the Education Testing Service for the Verbal Reasoning and Quantitative Reasoning measures of the computer-delivered GRE® General Test (GRE) (K. Haag, personal communication, October 14, 2016). In this particular testing program, they use number-correct scores to make routing decisions so there is no specific concern related to bias in provisional latent trait estimates. However, if latent trait estimates are used in a ca-MST to make routing decisions (e.g., information-based methods such as maximum Fisher information), the provisional ability estimates may be inaccurate in the presence of missing item data within a module. In such a case, the adaptive algorithms may not select appropriate modules for subsequent test stages, which may in turn compromise final, reported ability estimates. To summarize, missing responses within a module may affect the path an examinee takes during the test and, hence, the final estimate of the examinee's latent trait.

There are many studies that explored missing data in the context of item response theory modeling (De Ayala, Plake, & Impara, 2001; Eekhout et al., 2014; Finch, 2008; Sulis & Porcu, 2017; Van Ginkel, Sijtsma, van der Ark, & Vermunt, 2010). However, to our knowledge, there is no study that has investigated the impact of missing responses and how to deal with missingness in ca-MST environments. Given the increased interest in ca-MST administration and the problems that missing data may introduce for such test administration, it is imperative to explore methods for remediating such problems.

The aim of this study is to investigate ability estimation in ca-MST in the presence of missing data. We explore a variety of different methods for handling missing data when such data present itself under various missing data mechanisms, various magnitudes of missing responses, various test lengths, and various ca-MST designs. The methods examined in this study include treating missing responses as incorrect using person mean imputation, predicted probability imputation, and the full information

maximum likelihood (FIML) estimator on the full response matrix with missing values. In the remainder of the article, we present an overview of possible missing data methods for ca-MST and a Monte Carlo simulation study that shows how these methods performed under the conditions of our study.

An Overview of Missing Data Methods

Since missing data can threaten the accuracy of analysis results in various situations, searching for methods to handle missing values adequately has been an ongoing endeavor. Briefly, general missing data handling methods fall under four major umbrellas: (a) deleting cases with missing values (i.e., listwise deletion), (b) using the full information maximum likelihood (a.k.a. available case analysis or direct likelihood), (c) using single imputation, and (d) using multiple imputation (Rubin, 1987). Imputation methods replace the missing values with generated values from the data; on the other hand, the FIML method uses only available observed data without deletion or imputation (Kadengye, Ceulemans, & Van den Noortgate, 2013). Within psychometric literature, there are also modeling approaches built to incorporate (or tease out) missing item responses from the estimation of latent traits (e.g., Holman & Glas, 2005; Moustaki & Knott, 2000; Rose, von Davier, & Xu, 2010).

Deciding which missing data handling method might work well for the data challenge at hand depends on several aspects such as missingness mechanisms and data structure (Eekhout et al., 2014; Finch, 2008). There are three missing data mechanisms with different assumptions: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Based on Little and Rubin's (2002) descriptions, MCAR assumes that missing values are random and not stemming from a systematic cause. MAR assumes that missing values depend on other observed values in the data. MNAR assumes that missing values depend on either the variable itself or variables that are unobserved in the data.

Unfortunately, methods for handling missingness in item response data are sensitive to missing mechanisms. For example, listwise deletion results in unbiased estimates only if missingness is MCAR (Van Ginkel et al., 2010). Furthermore, the FIML and imputation methods are appropriate for MAR missingness (Gottschall, West, & Enders, 2012). Han and Guo (2014) argued that the missingness in computerized adaptive testing is often seen as MAR since the item selection process depends on test takers' observed performance on previous items. Similarly, Mislevy (2017) pointed out that the missingness mechanism in adaptive testing is MAR. However, these studies are in reference to item data that are missing due to the adaptive test not presenting an item to an examinee. In situations of adaptive or nonadaptive tests where examinees are permitted to skip items, item omissions may depend on the examinees' underlying ability levels, which can be considered MNAR (Wang, Jiao, & Xiang, 2013; Wolkowitz & Skorupski, 2013). Given that the missing data mechanism of skipped items can theoretically take on any form in a specific testing

context, we investigate in this study missing data methods in the presence of MCAR, MAR, and MNAR missingness in the ca-MST context.

Missing data methods may handle missingness by using an examinee's responses to other items and/or using the responses of other examinees (e.g., Finch, 2008; Sulis & Porcu, 2017). However, by the nature of ca-MST, data consist of an examinee's response vector in a particular module, and do not include responses of other examinees in the real-time assessment required to implement on-the-fly algorithms and reporting. Due to this reason, we focus our study on two single-imputation methods (i.e., PMI and PPI) and the FIML method, both of which are possible to use in the ca-MST real-time data context. Prior studies examined performances of these methods in other contexts (e.g., Finch, 2008). Although these methods are not an exhaustive set, they are selected as being common methods that would fit the data structure of ca-MST.

In PMI, the mean score of the observed responses of the person is calculated, and the missing values are replaced with values that are generated from random distributions defined by the mean score (Kadengye, Cools, Ceulemans, & Van den Noortgate, 2012). More specifically, the arithmetic mean of an examinee's observed scores is computed as follows:

$$PM_p = \frac{1}{I_{rp}} \sum_{i=1}^{I_{rp}} Y_{pi}, \quad (1)$$

where PM_p is the mean of the observed scores for person p , I_{rp} is the total number of item responses r for person p , and Y_{pi} is the observed response of person p for item i (Kadengye et al., 2012). To add random noise into imputation, PM_p is used as a probability to make a random draw from the Bernoulli distribution (Sijtsma & van der Ark, 2003). PPI adapts a logistic regression approach (a.k.a. proportional odds model; Eekhout et al., 2014). The log odds of the probability of correct response is obtained from the observed responses of an individual (van Buuren, 2010). This involves several steps: First, a logistic regression model is fitted such that

$$\ln\left(\frac{P(Y_i = 1)}{P(Y_i = 0)}\right) = a + Y_{-i}\beta, \quad (2)$$

where probability $P(Y_i = 1)$ is predicted from the responses on the present data in the module, Y_{-i} is the set of all observed responses except item i , and a and β are estimated parameters. Second, for each incomplete response, the predicted probability is calculated as

$$w_{na} = \exp(\hat{a} + Y_{-i}\hat{\beta}) / (1 + \exp(\hat{a} + Y_{-i}\hat{\beta})), \quad (3)$$

where na represents the data code for missing responses. Finally, random values are drawn from $u_i = \text{unif}(0,1)$. If $u_i > w_{na}$, missing values are imputed as 0, otherwise as 1.

The FIML method utilizes all available cases to directly estimate parameters (Peyre, Leplege, & Coste, 2011). It is important to recall that missing values are neither filled nor deleted in this method. Briefly and as discussed in Enders (2001), in the FIML method person likelihood (L_p) of the observed data is obtained by maximizing

$$\log L_p = K_p - \frac{1}{2} \log |\sum_p| - \frac{1}{2} (Y_p - \mu)' \sum_p^{-1} (Y_p - \mu), \quad (4)$$

where Y_p is the vector of complete data for person p , μ contains the corresponding mean estimates, K_p is a constant that depends on the number of complete-data points for person p , and $\sum_p$ is the model-implied covariance matrix.

Finally, the imputation method of treating missing values as incorrect is conducted by replacing the missing values with zero because zero indicates an incorrect response in binary response patterns. This method may be shown as

$$Y_{pi} = \begin{cases} 1 & \text{if } Y_{pi} = 1, \\ 0 & \text{if } Y_{pi} = 0, \\ 0 & \text{if } Y_{pi} = \text{NA}. \end{cases} \quad (5)$$

Method

Design Overview

This simulation study examined the performance of the four missing data handling methods (PMI, PPI, FIML, and incorrect) described in the previous section, within the context of a variety of testing conditions that may occur in ca-MST administrations. Specifically, the four imputation approaches were tested across a ca-MST design factor with four levels (1-3, 1-2-2, 1-2-3, and 1-3-3 designs), a test length factor with two levels (30-item and 60-item), a missingness-type factor with three levels (MCAR, MAR, and MNAR), and a percentage of missingness factor with four levels (5%, 15%, 30%, and 50%). All manipulated conditions were fully crossed, resulting in 384 conditions. In addition, complete data (e.g., no missingness conditions) results were used as a baseline condition. One hundred replications were performed within each condition. To simulate examinees, 900 theta values were generated from a normal distribution, $N(0, 1)$. All simulations were completed using R (R Development Core Team, 2016).

Ca-MST Simulations

The item parameters were based on a previous study (Schnipke & Reese, 1999). The item parameters and number of items in the different modules are given in Table 1. We generated three nonoverlapping essentially parallel ca-MST panels to hold the maximum panel exposure rates at 1/3. After the panels were built, we randomly

Table 1. Item Parameters and Number of Items in the Modules

Variable	Routing		Easy items		Medium items		Hard items	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
<i>a</i>	0.75	0.25	0.90	0.25	0.90	0.25	0.90	0.25
<i>b</i>	0.0	0.80	−1.0	0.50	0.0	0.50	1.0	0.50
<i>c</i>	0.15	0.25	0.15	0.25	0.15	0.25	0.15	0.25
Number of items	663		210		221		206	

Note. *a* = Discrimination; *b* = Difficulty; *c* = Guessing; *M* = mean, *SD* = standard deviation.

assigned 900 examinees to the panels with 300 examinees in each panel. We used IBM CPLEX program (ILOG, 2006) to build modules. When constructing the modules, items were obtained from the item pool with three fixed theta scores, $\theta_1 = -1$, $\theta_2 = 0$, $\theta_3 = 1$, which represent the target information function for easy, medium, and hard modules, respectively, as done in Chen, Chang, and Wu (2012). This means that the target information function is peaked around -1 , 0 , and 1 theta points for the item groups in easy, medium, and hard modules, respectively. The panel-level test information functions across the panel design and test length conditions are given in Figure 1. It is important to note that in the 1-2-2 design, there was no medium module in stages two and three, and in the 1-2-3 design there was no medium module in stage two. The routing module items were chosen from medium difficulty items (i.e., items that maximize the information function at theta point of 0). In the two-stage design (e.g., 1-3), there were 15 items in each module when the test length was 30. In the three-stage designs (e.g., 1-2-2, 1-2-3, and 1-3-3), there were 10 items in each module when the total test length was 30. Under the 60-item conditions, item numbers were doubled in each stage. In all ca-MST simulation conditions, the maximum Fisher information method (Lord, 1980) was used as the routing method. This method selects the next module as the one that maximizes information at one's provisional ability estimate. The expected a posteriori method (Bock & Mislevy, 1982) with a priori distribution of $N(0, 1)$ was used for provisional and final ability estimations.

Missing Data Generation

Data generation occurred in two steps. First, complete data were generated, and then the missing values were embedded into the data sets. The three parameter model (3PL) was chosen to calculate the probability of correct responses for data generation (Birnbaum, 1968). The 3PL model defines the conditional probability as

$$P(X_{ip} = 1 | \theta_p, b_i, a_i, c_i) = c_i + (1 - c_i) \frac{\exp[a_i(\theta_p - b_i)]}{1 + \exp[a_i(\theta_p - b_i)]}, \tag{6}$$

Figure 1. Test information functions for all panels across ca-MST and test length conditions.

where b_i is the item difficulty parameter, a_i is the item discrimination, c_i is the item guessing parameter, and θ_p is the latent trait score for person p . Equation (6) was used to obtain the probability of the correct response given person and item parameters.

After generating the complete response patterns for each case, missing responses were obtained by deleting some of the values based on the assumptions of missingness mechanisms. For MCAR missingness, probability of missingness, P_{mis} , was dependent on percentage of missingness (Enders, 2004). Random values were drawn from $u_i = \text{runif}(0, 1)$, and if $P_{mis} > u_i$, the person observed response Y_p would be recoded as missing, or *NA*. Otherwise, the person response Y_p remained in the data.

MAR data missingness was generated by following the methodology as shown in Finch (2008). For MAR missingness, the probability of missingness was dependent on the observed responses on the other items, y_{-i} , not including the target item y_i . The random values were drawn from $u_i = \text{runif}(0, 1)$. If $P(y_{-i}) > u_i$, then the person observed response Y_p was recoded as missing, or *NA*. Otherwise, Y_p remained in the data.

For MNAR missingness, the probability of missingness was dependent on a person's theta value (Wang et al., 2013; Wolkowitz & Skorupski, 2013), such that

$$P(R_{ip} = 1 | \theta_p) = \frac{\exp[x_i - (\theta_p)]}{1 + \exp[x_i - (\theta_p)]}, \quad (7)$$

where R_{ip} is the proportion of missingness, x_i is the intercept, and γ is the effect of the theta on missingness. Logistic regression is used to generate missing values where the intercept (e.g., x_i) is obtained by taking the log of the intended percentage of missingness (i.e., $\log(0.05/0.95)$ for 5% missingness), and the coefficient value (e.g., γ) is

set to 0.8 as a strong effect of the theta on missingness. Equation (7) indicates that having missing responses increases as the individual theta values decrease.

In addition to the baseline condition where there was no missing values, we manipulated four different percentages for missingness, that is, 5%, 15%, 30%, and 50% (Enciso, 2016; Zhang & Walker, 2008). The percentage of missingness was generated for each module by taking into account the total sample in the module and, as a result, the number of missing responses for each person was different. This may in fact potentially reflect realistic testing conditions in which examinees have different amounts of missing data. Since maximum likelihood estimators do not produce person estimates associated with constant response vectors, we ensured in the simulation that the first item response in each module did not have missing values.

Evaluation Criteria

The results of the study were evaluated in terms of mean bias ($\bar{e}$), root mean square error (RMSE), and the correlation between estimated and true theta ($\rho_{\hat{\theta}\theta}$). Mean bias was calculated as

$$\bar{e} = \frac{\sum_{j=1}^N (\hat{\theta}_j - \theta_j)}{N}, \quad (8)$$

where $\hat{\theta}_j$ is estimated theta, θ_j is true theta, and N is the sample size. Based on the criteria of Hoogland and Boomsma (1998), values of relative mean bias less than .05 were considered unbiased. RMSE was calculated as

$$RMSE = \sqrt{\frac{\sum_{j=1}^N (\hat{\theta}_j - \theta_j)^2}{N}}. \quad (9)$$

The correlation between true and estimated theta values was calculated as

$$\rho_{\hat{\theta}, \theta_j} = \frac{cov(\hat{\theta}_j, \theta_j)}{\sigma_{\hat{\theta}_j} \sigma_{\theta_j}}, \quad (10)$$

where $\sigma_{\hat{\theta}_j}$ is the standard deviation of estimated theta, and σ_{θ_j} is the standard deviation of true theta. These three overall outcomes were calculated separately for each replication across the 900 examinees, and then averaged across 100 replications.

Results

Overall, the results were defined under four main subsections: baseline condition results (i.e., the results of complete data), bias of theta estimates, RMSE of theta estimates, and the correlation between estimated and true thetas.

Table 2. Results of Complete Data.

Design	Test length	Bias	RMSE	Correlation
I-3	30	0.005	0.381	0.936
I-3	60	−0.001	0.298	0.962
I-2-2	30	−0.015	0.397	0.932
I-2-2	60	0.008	0.310	0.960
I-2-3	30	0.013	0.375	0.938
I-2-3	60	0.002	0.297	0.962
I-3-3	30	−0.002	0.377	0.939
I-3-3	60	0.007	0.297	0.963

Note. RMSE = root mean square error.

Baseline Condition Results

Bias, RMSE, and the correlation between estimated and true theta values of complete data calibrations are presented in Table 2, with these complete data sets serving as baseline conditions for the remainder of the study results. The first finding from these baseline conditions was that, regardless of the ca-MST design and test length, no bias was observed in theta estimates. The second finding was that, regardless of the design, as the test length increased the bias and RMSE of theta estimates decreased, and the correlation between true and estimated theta increased. The third finding was that there was no meaningful difference between the ca-MST designs on the outcomes, when comparing results obtained from conditions with the same test length.

Mean Bias Results

Both the type of imputation and the percentage of missingness played important roles on the bias of theta estimates. Specifically, a main finding with respect to mean bias outcomes was that higher amounts of bias were obtained when missing values were imputed as incorrect responses, and this finding held across all percentages of missingness. We also found that bias increased as the percentage of missingness increased, regardless of the type of imputation. This finding was most obvious when missing values were imputed as incorrect scores. A graphical depiction of this situation is displayed in Figure 2, and the results across the imputation types and various percentages of missingness are presented in Table 3. There were small differences in the bias of theta estimates between the missing data handling methods (see Figure 2), but the performance of FIML was slightly better than the other conditions, especially under 30% and 50% of missingness (see Table 3).

Root Mean Square Error Results

Both the type of imputation and percentage of missingness played an important role on the RMSE estimates. The main finding of the RMSE results was that, regardless

Figure 2. Bias in theta estimation by imputation type and percentage of missingness.

Table 3. Mean Bias in Theta Estimation by Imputation Type and Percentage of Missingness.

PM	Incorrect	FIML	PMI	PPI
5	−0.154	0.001	−0.005	−0.005
15	−0.424	0.041	0.026	0.026
30	−1.025	0.000	−0.042	−0.042
50	−1.849	−0.116	−0.225	−0.224

Note. PM = percentage of missingness; Incorrect = treating missing values as incorrect response; FIML = full information maximum likelihood; PMI = person mean imputation; PPI = predicted probability imputation.

of the imputation type, RMSE increased as the percentage of missingness increased. A graphical depiction of this result is displayed in Figure 3, and the results across the imputation type and various percentages of missingness are presented in Table 4. The effect of the percentage of missingness was more pronounced when missing values were imputed as incorrect scores. Regardless of the percentage of missingness, the incorrect imputation approach was the least preferred method with respect to RMSE. Another finding was that regardless of the percentage of missingness, the FIML method always produced lower RMSE than the other methods, especially under conditions with 50% of missingness. Also, we observed that the RMSE findings were identical for the PMI and PPI imputation methods across conditions.

Correlation Results

The percentage of missingness, type of imputation, and missingness type played important roles in the observed correlations between true and estimated theta. The

Figure 3. RMSE of theta estimation by imputation type and percentage of missingness.

Table 4. RMSE in Theta Estimation by Imputation Type and Percentage of Missingness.

PM	Incorrect	FIML	PMI	PPI
5	0.394	0.351	0.361	0.360
15	0.587	0.380	0.413	0.413
30	1.148	0.412	0.508	0.508
50	1.959	0.542	0.806	0.806

Note. RMSE = root mean square error; PM = percentage of missingness; Incorrect = treating missing values as incorrect response; FIML = full information maximum likelihood; PMI = person mean imputation; PPI = predicted probability imputation.

main finding of this outcome was that, regardless of the missingness type (i.e., MAR, MCAR, or MNAR), the magnitude of the correlation between estimated and true thetas decreased as the percentage of missingness increased. This was consistent with the bias and RMSE findings. A graphical depiction of the relationships between this outcome and the factors of imputation type, percentage of missingness, and missingness type are displayed in Figure 4. The results of the correlation outcomes across the imputation type, percentages of missingness, and missingness type are presented in Table 5. Another finding was that, regardless of the missingness type and percentage of missingness, FIML produced the same or higher magnitude of correlations across the conditions with a few exceptions. The exceptions occurred for the incorrect imputation method under MNAR missingness type across all percentages of missingness (see Figure 4). We also found that when the missingness rate was 5% or 15%, all imputation methods resulted in similar magnitudes of correlations.

Figure 4. Correlation between estimated and true theta by imputation type, missingness type, and percentage of missingness.

However, as the percentage of missingness increased, the effect of imputation type on the correlation outcomes became more obvious. When the missingness was 50%, treating missing values as incorrect resulted in lower magnitudes of correlations, especially when the missingness type was MNAR. Conversely, there were no meaningful differences in the correlation outcomes between PMI and PPI imputation methods across all conditions.

Discussion

To date, many aspects of computerized adaptive testing have been studied, including comparisons of routing methods (Luecht, 2000), proposal of new test assembly (Luecht & Nungester, 1998), specifications of stages and modules (Patsula, 1999; Zenisky, 2004), and issues of score equating (Davey & Lee, 2011). However, no study has examined issues of missing data in ca-MST. This study combines missing data and ca-MST contexts and provides a unique insight into the ability estimation in ca-MST in the presence of missing responses. We approached this study under the assumption that missing data is a common occurrence in testing and almost inevitable under some particular test administration methods. This study contributed to the literature on missing data in ca-MST and, more specifically, aimed to show the

Table 5. Correlation Between Estimated and True Theta by Imputation Type, Percentage of Missingness, and Missingness Type.

PM	MT	Correlation			
		Incorrect	FIML	PMI	PPI
5	MAR	0.941	0.946	0.942	0.942
5	MCAR	0.935	0.945	0.942	0.942
5	MNAR	0.950	0.945	0.943	0.943
15	MAR	0.924	0.938	0.925	0.925
15	MCAR	0.902	0.938	0.927	0.927
15	MNAR	0.950	0.934	0.924	0.924
30	MAR	0.893	0.922	0.888	0.889
30	MCAR	0.817	0.922	0.893	0.892
30	MNAR	0.942	0.912	0.882	0.882
50	MAR	0.851	0.886	0.800	0.800
50	MCAR	0.695	0.891	0.830	0.829
50	MNAR	0.910	0.874	0.796	0.796

Note. PM = percentage of missingness; MT = missingness type; Incorrect = treating missing values as incorrect response; FIML = full information maximum likelihood; PMI = person mean imputation; PPI = predicted probability imputation; MCAR = missing completely at random; MAR = missing at random; MNAR= missing not at random.

relative strengths and weaknesses of imputation and estimation methods in ca-MST with respect to ability estimation.

Treating missing values as incorrect may be considered the simplest practical solution to dealing with missing values in ca-MST. However, as stated in De Ayala et al. (2001), treating missing values as incorrect has been shown to produce bias in theta estimates. In addition, Mislevy (2017) pointed out that treating missing test responses as incorrect may result in negatively biased theta estimates in a variety of testing conditions. Therefore, we evaluated multiple alternative methods for researchers and practitioners in dealing with missing values in ca-MST administrations. This study investigated how accurately ability could be estimated when using different missing data handling approaches rather than treating missing values as incorrect. Aligning with the literature, this study found that treating missing values as incorrect produced biased theta estimates in ca-MST environment.

The findings of this study revealed that using FIML, PMI, and PPI recovered ability estimates well according to multiple evaluation criteria. Compared to imputing the data as incorrect responses, these methods obtained lower amounts of bias, lower RMSE results, and higher correlations between true and estimated ability, no matter the type of missingness. In fact, when the percentage of missingness was 5%, the results of these three missing data methods were very similar to the baseline condition results (e.g., complete data). Even when the percentage of missingness was 15% or 30%, the three alternative missing data handling methods resulted in acceptable theta estimates, regardless of the missingness type, the MST design, or the test length.

However, in conditions where 50% of the item data was missing, we always obtained biased ability estimate results no matter the method for handling missing data. This aligns with previous studies that found missing data methods to perform better in the presence of less missing data (see Enders, 2004; Finch, 2008; Sijtsma & van der Ark, 2003). In addition, this study found that neither the length of the test nor the type of ca-MST design played important roles in the outcomes.

This study does not argue against treating missing values as incorrect. It is important to note that treating missing values as incorrect could be the best choice when using the number-correct score routing method. This is because in this routing approach, the number of right answers is taken into account when selecting the subsequent modules and skipping an item may be considered the same as being unable to answer an item correctly. Zhang and Walker (2008) mentioned that treating missing values as incorrect may also perform better in testing situations where omitted responses are mostly observed for more difficult items in paper-and-pencil tests. Also, it may be that informing examinees that skipped items will be scored as incorrect can assist in deterring item skipping. However, in ca-MST, modules are selected based on the ability estimates of examinees from the previous module, and hence there may be testing situations where scoring items as incorrect can have major ramifications for module selection, and possibly negative ramifications. In such scenarios, our study demonstrated that alternative forms of missing data handling may be worthy of consideration.

Some studies have recommended against using mean imputation to handle missing data (e.g., Eekhout et al., 2014). However, we found that in ca-MST settings PMI appeared to perform similarly to the other studied imputation methods and to outperform the practice of imputing missing data as incorrect. In addition, Peyre et al. (2011) found that personal mean score imputation may perform better than multiple imputation and the FIML with descriptive parameters. Hence, even though some studies argued that PMI is outperformed by other missing data methods, adding random noise to PMI might have helped to increase the performance of PMI in our ca-MST simulation.

Some studies have discussed the use of FIML as a practical choice for handling missing data, since it does not require additional processing for missing values and it is often found to have similar performance to some imputation approaches (Kadengye et al., 2012; Kadengye et al., 2013; Zhang & Walker, 2008). Aligning with these studies, we concluded that the FIML approach performed as well as the other methods in terms of bias of the ability estimates. Additionally, it provided the lowest RMSE in many conditions and produced the most consistent results across all of the factors in this study. For example, the percentage of missingness was one of the most influential factors across all conditions in our study, yet results from the use of FIML were least affected by differences in such percentages (see Figures 2 and 3). This was an important finding because one never knows the percentage of missingness for a person before a test administration. Therefore, for researchers and

practitioners whose goal is to obtain unbiased ability estimation in ca-MST, FIML might be one of the best available approaches for dealing with missingness.

This study is limited by the use of single imputation methods to replace missing values. For future research, this study might be extended to multiple imputations of PPI and PMI by replacing missing values multiple times, analyzing the multiple sets of data, and pooling the results. In addition, this study did not take into account the time limit in adaptive testing. When a time limit exists some test takers may not answer all questions, which results in missingness in the not-reached items. Future studies may aim to investigate methods to deal with missingness of not-reached items in the context of ca-MST. As a final limitation, it is important to note that dealing with missing data is often just as much of a practical decision as it is a statistical one. For example, when the method to deal with missing responses is known by the test-takers, the test-takers may develop test taking strategies based on these missing data handling methods. Such strategies are expected to impact ability estimates but are not taken into consideration in our simulation. There is definitely a need for future studies that investigate ca-MST missing data from the test takers' perspectives.

Declaration of Conflicting Interests

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

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.
- Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6, 431-444.
- Brandriet, A., & Holme, T. (2015). Methods for addressing missing data with applications from ACS exams. *Journal of Chemical Education*, 92, 2045-2053.
- Chen, P.-H., Chang, H.-H., & Wu, H. (2012). Item selection for the development of parallel forms from an IRT-based seed test using a sampling and classification approach. *Educational and Psychological Measurement*, 72, 933-953.
- Cohen, J. (1973). Eta squared and partial eta-squared in fixed factor ANOVA designs. *Educational and Psychological Measurement*, 33, 107-112.
- Davey, T., & Lee, Y.-H. (2011, June). *Potential impact of context effects on the scoring and equating of the multistage GRE® Revised general test* (Research Report RR-11-26). Princeton, NJ: Educational Testing Service. Available from <https://files.eric.ed.gov/fulltext/EJ1110374.pdf>

- De Ayala, R. J., Plake, B. S., & Impara, J. C. (2001). The impact of omitted responses on the accuracy of ability estimation in item response theory. *Journal of Educational Measurement*, 38, 213-234.
- DeMars, C. (2002). Incomplete data and item parameter estimates under JMLE and MML estimation. *Applied Measurement in Education*, 15, 15-31.
- Eekhout, I., de Vet, H. C., Twisk, J. W., Brand, J. P., de Boer, M. R., & Heymans, M. W. (2014). Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. *Journal of Clinical Epidemiology*, 67, 335-342.
- Enciso, S. M. S. (2016). *The effects of missing data treatment on person ability estimates using IRT models* (Unpublished master's thesis). University of Nebraska-Lincoln, Nebraska.
- Enders, C. K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling*, 8, 128-141.
- Enders, C. K. (2004). The impact of missing data on sample reliability estimates: Implications for reliability reporting practices. *Educational and Psychological Measurement*, 64, 419-436.
- Finch, H. (2008). Estimation of item responses theory parameters in the presence of missing data. *Journal of Educational Measurement*, 45, 225-245.
- Gottschall, A. C., West, S. G., & Enders, C. K. (2012). A comparison of item level and scale level multiple imputation for questionnaire batteries. *Multivariate Behavioral Research*, 47(1), 1-25.
- Han, K. T., & Guo, F. (2014). Impact of violation of the missing at random assumption on full information maximum likelihood method in multidimensional adaptive testing. *Practical, Assessment, Research and Evaluation*, 19(20). Retrieved from <https://pareonline.net/getvn.asp?v=19&n=2>
- Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and meta-analysis. *Sociological Methods and Research*, 26, 329-367.
- Holman, R., & Glas, C. A. W. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical & Statistical Psychology*, 58, 1-17.
- ILOG. (2006). *ILOG CPLEX 10.0* (User's manual). Paris, France: Author.
- Kadengye, D. T., Ceulemans, E., & Van den Noortgate, W. (2013). Direct likelihood analysis and multiple imputation for missing item scores in multilevel cross-classification educational data. *Applied Psychological Measurement*, 38, 61-80.
- Kadengye, D. T., Cools, W., Ceulemans, E., & van den Noortgate, W. (2012). Simple imputation methods versus direct likelihood analysis for missing item scores in multilevel educational data. *Behavioral Research*, 44, 516-531.
- Little, R. J. A., & Rubin, D. (2002). *Statistical analysis with missing data*. New York, NY: John Wiley.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Luecht, R. M. (2000, April). *Implementing the computer-adaptive sequential testing (CAST) framework to mass produce high quality computer-adaptive and mastery tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Luecht, R. M., Brumfield T., & Breithaupt, K. (2006). A testlet assembly design for adaptive multistage tests. *Applied Measurement in Education*, 19, 189-202.

- Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement*, 35, 229-249.
- Maher, J. M., Markey, J. C., & Ebert-May, D. (2013). The other half of the story: Effect size analysis in quantitative research. *Life Science Education*, 12, 345-351.
- Mislevy, R. J. (2017). *Missing responses in item response modeling*.
- Moustaki, I., & Knott, M. (2000). Weighting for item non-response in attitude scales by using latent variable models with covariates. *Journal of the Royal Statistical Society*, 163, 445-459.
- Patsula, L. N. (1999). *A comparison of computerized adaptive testing and multistage testing*. Available from ProQuest Dissertations & Theses Global. (Order No. 9950199)
- Peyre, H., Leplege, A., & Coste, J. (2011). Missing data methods for dealing with missing items in quality of life questionnaires. A comparison by simulation of personal mean score, full information maximum likelihood, multiple imputation, and hot deck techniques applied to the SF-36 in the French 2003 decennial health survey. *Quality of Life Research*, 20, 287-300.
- R Development Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.rproject.org>.
- Rose, N., von Davier, M., & Xu, X. (2010). *Modeling nonignorable missing data with item response theory* (ETS Research Report No. RR-10-11). Princeton, NJ: Educational Testing Service.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley.
- Schnipke, D. L., & Reese, L. M. (1999). *A comparison of testlet-based test designs for computerized adaptive testing* (Law School Admission Council Computerized Testing Report). Newtown, PA: Law School Admission Council.
- Sijsma, K., & van der Ark, L. A. (2003). Investigation and treatment of missing items scores in test and questionnaire data. *Multivariate Behavioral Research*, 38, 505-528.
- Sulis, I., & Porcu, M. (2017). Handling missing data in item response theory. Assessing the accuracy of a multiple imputation procedure based on latent class analysis. *Journal of Classification*, 34, 327-359.
- van Buuren, S. (2010). Item imputation without specifying scale structure. *Methodology*, 6(1), 31-36.
- Van Ginkel, J. R., Sijsma, K., van der Ark, L. A., & Vermunt, J. K. (2010). Incidence of missing item scores in personality measurement, and simple item score imputation. *Methodology*, 6(1), 17-30.
- Wang, S., Jiao, H., & Xiang, Y. (2013, April). *The effect of nonignorable missing data in computerized adaptive test on item fit statistics polytomous item response models*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
- Wolkowitz, A. A., & Skorupski, W. P. (2013). A method of imputing response options for missing data on multiple choice assessments. *Educational and Psychological Measurement*, 73, 1036-1053.
- Zhang, B., & Walker, C. M. (2008). Impact of missing data on person model fit and person trait estimation. *Applied Psychological Measurement*, 32, 466-479.
- Zenisky, A. L. (2004). *Evaluating the effects of several multi-stage testing design variables on selected psychometric outcomes for certification and licensure assessment*. Available from ProQuest Dissertations & Theses Global. (Order No. 3136800)