ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2022; 13(3); 212-231

An Investigation of the Effect of Missing Data on Differential Item Functioning in Mixed Type Tests*

Leyla Burcu DİNÇSOY** Hülya KELECİOĞLU***

Abstract

In this research, the aim was to examine the effects of Markov Chain Monte Carlo (MCMC), multiple imputation (MI), and expectation maximization (EM), all methods of coping with missing data in mixed type tests including dichotomous and polytomous items, on the differential item functioning (DIF). The study was carried out on a complete data set consisting of the scores of 1160 students who took booklet number 9 in the science test in Trends in International Mathematics and Science Study (TIMSS) 2019 and answered it in full. The conditions to be examined for the effectiveness of the methods were missing data mechanism (MCAR and MAR), DIF level (A, B, and C), and missing data rate (10% and 20%). Data were assigned to the missing data sets created by deleting data at different rates under the missing completely at random (MCAR) and missing at random (MAR) mechanisms over the aforementioned data set. DIF analysis was performed on all the data sets obtained with the poly-SIBTEST method using the MCMC, MI, and EM methods. The results obtained from the complete data set were then compared with the result implications from other data sets for reference. The study showed that the EM and MCMC methods performed better for the C-level DIF than the A and B levels in terms of all conditions examined. MI was observed to be the most successful method in determining DIF in items showing DIF in 10% and 20% MCAR mechanisms. When compared with the complete data set, the three methods showed similar results in the 10% MAR mechanism while MCMC gave the closest results in the 20% MAR mechanism.

Keywords: mixed type test, missing data, differential item functioning, poly-SIBTEST

Introduction

The need to evaluate many individuals in a short period of time led to the development of achievement and aptitude tests in education (Özgüven, 2017). The assessments made as a result of these tests need to allow valid interpretations because they have individual, social, and political consequences. At this point, the problems of bias and missing data, which may lead to mistaken interpretations of the test results gain in importance.

It is essential that such achievement and aptitude tests do not contain bias and that they are fair for all those being tested. The technique known as DIF analysis is used in to statistically process bias (Zumbo, 1999). DIF refers to the differing performances of students with the same ability on an item. For an item that does not show DIF, it is expected that individuals with the same ability levels have the same probability of responding to the item correctly even if the individuals belong to different groups. However, if different item difficulties are observed in different groups with the same ability levels, the item exhibits DIF (Millsap & Everson, 1993). Camili (2006) examined DIF determination methods in two different groups: methods that equalize individuals by using observed scores and methods based on item response theory. Hambleton et al. (1993) divided DIF methods into three categories: those based on classical test theory, those based on item response theory, and those based on chi-square. Different researchers have classified DIF detection methods in different ways. Methods based on the classical test theory include Mantel-Haenszel (MH), logistical regression (LR), and the standardization method. Meanwhile, methods based on item response theory include Lord's chi-square, the likelihood ratio, and

* This study is a part of master's thesis conducted under the supervision of Prof. Dr. Hülya KELECİOĞLU and prepared by Leyla Burcu DİNÇSOY.

** Teacher, Ministry of National Education, Ankara-Turkey, leylaburcuadak@gmail.com, ORCID ID: 0000-0002-5633-3520
*** Prof. Dr., Hacettepe University, Faculty of Education, Ankara-Turkey, hulyakelecioglu@gmail.com, ORCID ID: 0000-0002-0741-9934

To cite this article:
Dinçsoy, L. B., & Kelecioğlu, H. (2022). An investigation of the effect of missing data on differential item functioning in mixed type tests. *Journal of Measurement and Evaluation in Education and Psychology, 13*(3), 212-231. https://doi.org/10.21031/epod.1091085

Received: 21.03.2022
Accepted: 16.09.2022

Raju's area measurements (Camili, 2006; Zumbo, 1999). Many methods originally developed for items scored in two categories have later been expanded to include items scored in multiple categories. Some of these methods are MH, LR, and SIBTEST (Swaminathan & Rogers, 1990). This study used Poly-SIBTEST, the generalized version of the SIBTEST method that can be used for two- and multi-category data.

DIF may adversely affect the reliability of test scores as it may result in erroneous assessments for test takers. However, DIF is not the only factor that can lead to the increased validity of test scores. For instance, missing data is another such factor. It is also possible for both the DIF and the missing data to occur simultaneously (Garret, 2009). A review of the literature reveals studies whose primary aim was to determine the methods of coping with missing data that show better performance under different conditions (such as sample size, focus-reference group ratio, missing data rate, missing data mechanisms, polytomous-dichotomous items, mixed type tests, etc.), (Banks & Walker, 2006; Emenogu et al., 2010; Falenchuk & Herbert, 2009; Finch, 2011a; Garrett, 2009; Nichols et al., 2022; Sedivy et al., 2006; Tamcı, 2018) or the aim was to compare the performances of DIF detection methods in the presence of missing data (Finch, 2011b; Garrett, 2009; Robitzsch & Rupp, 2009; Rousseau et al., 2006; Sedivy et al., 2006). For example, Finch (2011a) examined the effectiveness of the DMF method in the presence of missing data in the uniform DMF analysis by calculating type 1 error and power ratios. He worked with items scored in two categories. He determined three different sample sizes and kept the focal reference group ratio constant. Under TRK, RK and ROK mechanisms, 5% and 10% of missing data were generated. It compared the effectiveness of the DMF, list-based deletion and zero assignment methods with MH, LR and SIBTEST methods by determining the DMF. He stated that the type 1 error rates for zero assignment are inflated in the RK mechanism, while the results in the TRK and ROK mechanisms are similar to the full dataset results. It is also stated that the type 1 errors and power ratios of the list-based deletion and RTA methods show similar results to the full dataset. Selvi and Alıcı (2018) examined the effect of missing data assignment methods on different DMF detection methods. The test consisting of eighty multiple-choice items was scored in two categories. In their study, BM and regression assignment were used as missing data handling methods, and MH, standardized method and likelihood ratio test were used as DMF detection methods. It was reported that the missing data assignment methods caused a difference in DMF items and this difference was significant in the MH method.

Deficiencies in the information collected for assessment cause a decrease in reliability and validity and increase the probability of inaccurate decisions (Turgut & Baykul, 2012). Missing data can occur for many different reasons. For instance, participants may deliberately choose not to respond to an item, overlook the item, forget to return to a skipped item, avoid answering the item, or not know the answer. Alternatively, the interviewer may skip the item, or it may not be suitable for the participant. The participant may even have to leave the study, or ultimately, errors might be made in the data entry stage (Allison, 2002). The assumption of almost all statistical methods is that all participants have complete information for the variables to be included in the analysis (Allison, 2002). For example, DIF detection methods such as Mantel-Haenszel (Holland & Thayer, 1988), logistic regression (Swaminathan & Rogers, 1990), and simultaneous item bias test or SIBTEST (Shealy & Stout, 1993) are not designed for datasets with missing data. When the solution is the exclusion of participants with missing responses from the analysis, a large reduction in sample size may result, and detection power may be restricted if DIF is present (Banks, 2015). Instead, there are methods that researchers can choose to impute values to replace missing data. These methods have a significant impact on statistical results (Garret, 2009). Incorrect method selection can be a source of bias as it may result in masking the actual DIF or generating incorrect DIF in items that are not actually DIF (Banks, 2015). In the presence of missing data in the data set, the missing data mechanisms should be examined. The missing data mechanism is the mathematical relationship between the variables and the probability of the data being missing (Enders, 2010). Missing data mechanisms generally fall into three categories, classified by Little and Rubin (2020) as missing completely at random (MCAR), missing at random (MAR), and missing at not random (MNAR). This classification is the most widely accepted. In the case that a Y variable has missing data, in order to be able to state that the missing data in the Y variable are in the MCAR mechanism, the probability of missing data in the Y variable should be unrelated to the Y variable and

other variables. The MCAR mechanism is when the probability of missing data in the Y variable is irrelevant to the Y value when all other variables in the analysis are controlled (Allison, 2002). The ROK mechanism, on the other hand, states that the probability of missing data in the Y variable is related to the Y variable even after other variables in the dataset are controlled. The probability of missing data depends on the missing variable (Enders, 2010). Alpar (2021) summarized these mechanisms with an example as follows: In a study where weight was examined with the gender variable, it might be said that the data is in the MCAR mechanism if there is no reason among all the participants who did not state their weight, the MAR mechanism if the rate of women not answering their weight is higher, and the MNAR mechanism if those with more or less weight did not answer their weight.

In this way, missing data encountered in different mechanisms posed problems in analysis, which led researchers to search for solutions. This search for solutions started in the 1930s but became popularized with the work of Rubin (1976) (Toka, 2012). In general, methods of missing data imputation are grouped under two categories: deletion and simple imputation and probabilistic and offset data imputation. Probabilistic and translational methods are further divided into two groups as those based on the maximum likelihood approach and the MI approach (Demir, 2013). Examples of methods based on multiple data imputation approaches include listwise deletion, pairwise deletion, mean value imputation to deletion, and simple assignment-based methods; EM algorithm to methods based on maximum likelihood approach, direct maximum likelihood, and Bayesian data imputation methods; and MI, random imputation, and MCMC methods. MCMC, one of the methods used in this study, produces chains where each of the simulated values is lower than the previous value, unlike standard Monte Carlo methods, which generate a series of independent values through a simulation from the desired probability distribution. A Markov chain is a stochastic process with the property that any value in the value imputation sequence depends only on the previous value in its chain, thus being independent of all other prior states. The basic principle of MCMC is that when this Markov chain goes through a sufficient number of iterations, it will reach the desired posterior distribution (Gill, 2002). EM, on the other hand, is an iterative method that makes maximum probability estimations in two steps: expectation and maximization. This method begins first with the estimation of the mean vector and the covariance matrix. To estimate the missing data from the variables observed in the expectation step, a set of regression equations is set up using the mean vector and covariance matrix. By means of the established regression equations, a value is assigned to the missing data (Enders, 2010). Another method, MI, is designed to make multiple imputations instead of assigning a single value to the missing data and creates more than one complete data set (Van Buuren, 2012). It is performed in three stages: assigning data m times for each missing data, applying standard analyzes with m completed data sets, and combining the obtained m analysis results (Alpar, 2021).

In attempts to eliminate the problems and biases caused by missing data through data imputation, unconsciously imputed data does not eliminate the problem and may damage the reliability of the results. (Çüm et al., 2018; Little & Rubin, 1987). Therefore, it is important to determine the effect of value imputation methods instead of missing data in the presence of DIF in order to reduce the possible threats of DIF and missing data on validity of the test.

Purpose of the Research

A large number of studies can be found that deal with imputing data or DIF determination instead of missing data, but studies that deal with both are limited. Examining the literature and focusing on the effect of missing data on differential item functioning reveals that most studies have been conducted on dichotomous simulation data (e.g., Banks & Walker, 2006; Emenogu, 2010; Falenchuck & Herbert, 2009; Finch, 2011a, 2011b; Nichols et al., 2022; Robitzsch & Rupp, 2009; Rousseau et al., 2006). There are also some studies in which polytomous data are used (e.g., Garrett, 2009; Sedivy et al., 2006), but there are few studies using real data (e.g., Raousseau, 2004; Selvi & Alıcı, 2018; Tamcı, 2018). Because these studies generally use simulated data, tests in polytomous and dichotomous items are used separately. Therefore, this study examined the effects of missing data imputation methods on DIF under different conditions on the real data set in a mixed type test containing both polytomous and

dichotomous items. Since mixed type tests are frequently encountered in practice, new studies with mixed type tests are necessary.

This study is essential and will contribute greatly to the literature because it has been carried out with mixed type tests and real data, and it examines the effects of imputing values on real mixed test data containing missing data at two different rates and in two different missing data mechanisms using different methods on the differential item functioning in comparison with complete data sets. It is also important to see the effectiveness of different methods selected in the different conditions determined in the study.

The DIF analysis process is affected by missing data, as is the case with many analyses. If there are missing values in the dataset, appropriate methods should be selected accordingly, and there an effort must be made to prevent any problems that may be caused by the missing data. In this case, it is important to evaluate how much DIF results are affected by the method used and how similar the results it provides are to the real situation. If the missing data are not successfully compensated for, an item with DIF may appear as being without DIF because of the imputation method or an item without DIF may show DIF. Likewise, changes may occur in the DIF levels of items with DIF.

In order for the tests to give valid and reliable results, it is important to use the missing data imputation method that best compensates for these situations. As stated by Banks and Walker (2006), Finch (2011a, 2011b), and Garrett (2009), it is important for researchers to use one of these methods since when appropriate value imputation methods are used, results similar to full data sets are usually obtained.

In this regard, the research questions addressed in this study are:

1. How are the DIF results obtained by imputing data by the MCMC, MI, and EM methods to the data sets created by deleting 10% and 20% data in accordance with the MCAR mechanism from the full data set obtained from the TIMSS 2019 Science test?

2. How are the DIF results obtained by imputing data by MCMC, MI and EM method to the data sets created by deleting 10% and 20% data in accordance with the MAR mechanism from the full data set obtained from the TIMSS 2019 Science test?

3. How is the distribution of DIF results obtained by imputing data with MCMC, MI and EM methods to the data sets created by deleting 10% and 20% data in accordance with MCAR and MAR mechanisms from the full data set obtained from the TIMSS 2019 Science test differ according to the items showing and not showing DIF?

Method

The Model of the Research

This study was carried out with a correlational study model to examine the effect of distinctive methods of coping with missing data on DIF using reference results obtained from complete datasets in different circumstances. While the survey method describes the existing situation, correlational studies examine how the variables are related to each other (Karasar, 2011).

Participants

The data used in the study were obtained from the responses of students who participated in the TIMSS 2019 study conducted by the International Association for the Evaluation of Educational Achievement (IEA). The population of the study, therefore, consisted of approximately 250 thousand students who participated in the eighth grade TIMSS 2019 assessment. The sample was made up of students who took booklet number 9 in the eighth grade TIMSS 2019 evaluation and from the five native or non-native English-speaking countries where science score averages are close to each other, and thus there can be

no source of DIF. All the booklets were examined, and the booklet number 9, which contains the highest number of polytomous items, was selected.

The student answers with missing data were removed, and 1160 students were included in the analysis. The distribution of the number of students in the data set according to the native language variable was examined as a source of DIF, and the science score averages of the countries and their native languages are provided in Table 1.

Table 1

Distribution of Students Who Took Booklet Number 9 in the TIMSS 2019 Science Test by Countries, Science Averages, and Languages of the Countries

Countries	Number of Students	Science Average*	Language
England	177	517(4.8)	English
America	403	522(4.7)	English
Sweden	192	521(3.2)	Swedish
Turkey	194	515(3.7)	Turkish
Portugal	194	519(2.9)	Portuguese

*Standard errors are given in parentheses ().

Data Collection Tools

The research data consisted of the responses given by students from England, America, Sweden, Portugal, and Turkey to the seventeen items in booklet number 9, where the polytomous items of the TIMSS 2019 science test were the highest. Twelve of the 17 items used in the booklet were multiple-choice, and five were open-ended. Open-ended items were polytomous items scored as 0-1-2. In order to limit the research, the sample size and focus-reference group ratio were kept constant in the study.

Table 2

Examined Conditions

Conditions	Levels
DIF Level	A
	B
	C
Missing Data Rate	%10
	%20
Missing Data Mechanism	MCAR
	MAR
Imputation Methods	MCMC
	EM
	MI (5 imputation)

Data Analysis

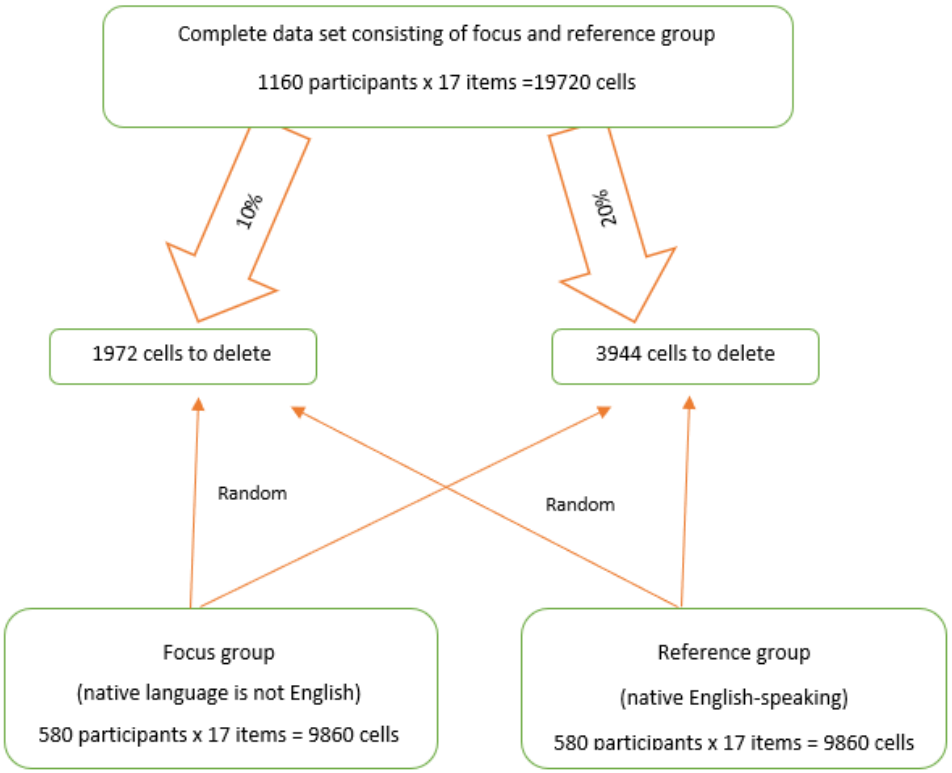
Creating Missing Datasets

Using the R program "missMethod" package from the complete dataset, four missing datasets were created in the MAR and MCAR mechanisms at rates of 10% and 20%. Missing data mechanisms constitute a bigger issue than the amount of missing data does. Though it is not an exact criterion, a missing rate of 5% or less with a random mechanism is negligible in large data sets (Tabachnick & Fidell, 1996). There are no strict rules regarding the negligibility of missing data with respect to their amounts. Similar studies in the literature show that these rates generally vary between 5% and 30%, and, in this study, the percentages were determined by taking these rates into account. The convenience of

the generated missing datasets for the missing data mechanism was examined using the IBM SPSS 24.0 program. The two datasets obtained in accordance with the MCAR mechanism were created by deleting an equal number of random data from each of the 17 items at rates of 10% and 20% from the complete dataset. The calculations related to the data deletion process in the MCAR mechanism are provided in Figure 1.

Figure 1

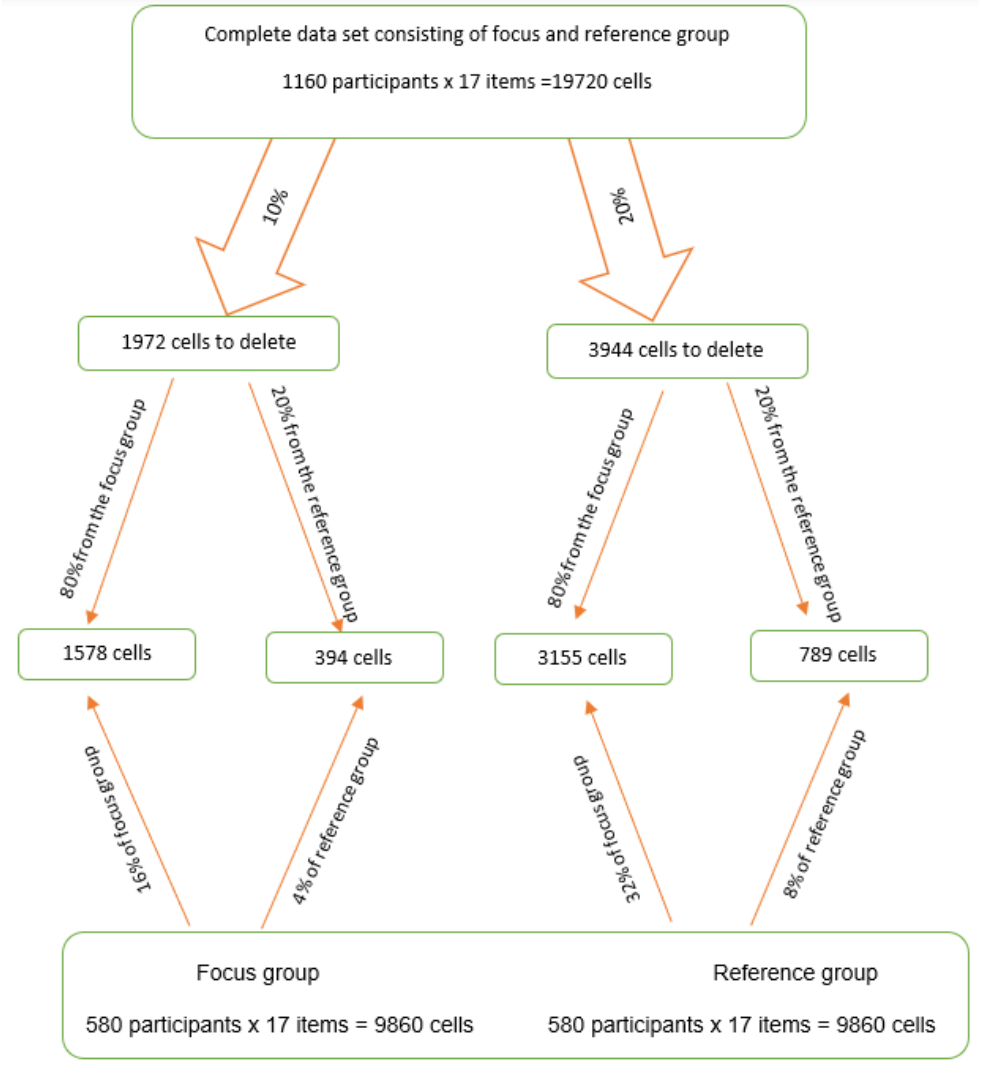
Deleting Missing Data Using the MCAR Mechanism



The process of generating missing data properly with the MAR mechanism was substantiated by deleting 80% of the data from the focus group and 20% of the data from the reference group randomly at rates of 10% and 20%. Calculations related to the data deletion in the MAR mechanism can be found in Figure 2.

Figure 2

Deleting Missing Data Using the MAR Mechanism



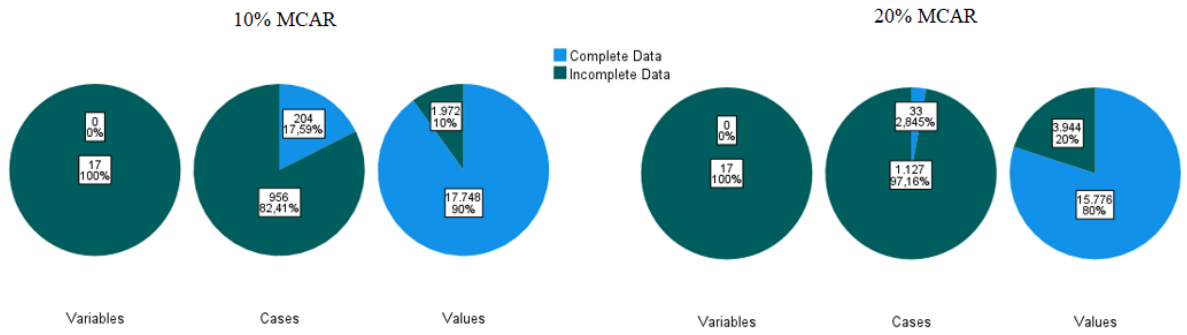
Examination of Missing Data Mechanisms

The examination of missing data mechanisms and imputing values instead of missing data were carried out on four datasets obtained at the previous stage. When the descriptive statistics of the datasets were examined before moving on to the missing data mechanism, no missing data for the native language variable was found in any datasets. In the two missing datasets created at 10%, 116 pieces of missing data were observed in each of the 17 items, a total of 1972 pieces of data. In the datasets created at the rate of 20%, 232 pieces of missing data were observed in each of the 17 items, and a total of 3944 pieces of data were examined.

The numbers and percentages of the missing data, which were determined by descriptive statistics in the variables, participants and values, are provided in the pie charts. Figure 3 shows the number of datasets with 10% and 20% missing data created by the MCAR mechanism in variables, participants, and values and their percentages.

Figure 3

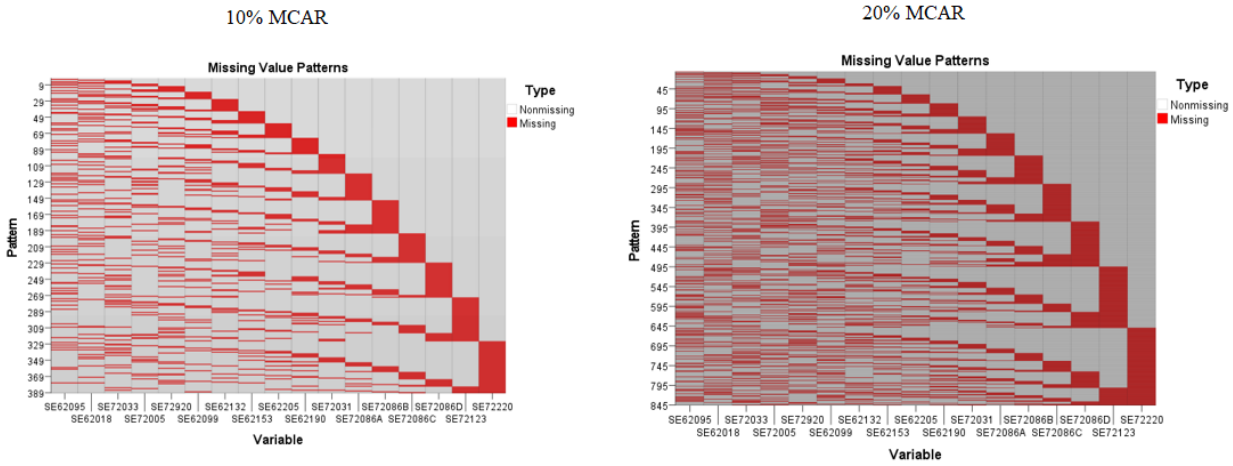
The ratio of Missing in the MCAR Dataset of 10% and 20% of the Data in Variables, Participants, and All Values



When Figure 3 is examined, it may be seen that there were missing data in all 17 variables (100%) in the variables graph in the 10% MCAR dataset, missing data were observed in 204 (17.59%) out of 1160 participants, and data were missing in 1972 (10%) of the 19720 cells in the data set. In the 20% MCAR data set, there were missing data in all 17 variables (100%) in the variables graph, missing data were observed in 33 (2.84%) out of 1160 participants, and data were missing in 3944 (20%) of the 19720 cells in the data set. Graphs showing missing data patterns are presented in Figure 4.

Figure 4

10% and 20% MCAR Dataset Missing Data Pattern Graphs



There are 389 and 845 patterns belonging to 17 variables and 1160 participants, respectively, at missing data rates of 10% and 20%. Since the gray cells representing the observed data are not clustered in the lower right part of the graph and the red cells representing the missing data are not clustered in the upper left part of the graph, it can be stated that the missing are not in any specific order and that there is a non-monotonous pattern.

Whether the missing in the missing datasets have the MCAR mechanism was also examined by the Little & Rubin (2020) MCAR test. Since the p values for both datasets were not statistically significant (p=0.099 for a 10% missing dataset, p=0.656 for a 20% missing dataset), this served as evidence that the data were completely random. Although there are statistical tests regarding the compatibility of the missing data for the MCAR mechanism, this is not the case for the MAR mechanism. In this case, characterizing the data as MAR is only an assumption (Schafer & Graham, 2002).

Imputation Methods Instead of Missing Data

The missing data were imputed to the created datasets using the MCMC, EM, and MI methods. The MCMC imputation was performed with Lisrel 8.80, and the imputations were performed with the EM and MI methods with the IBM SPSS 24.0 program. Eight datasets were generated by imputing four datasets with the rates of 10% and 20% missing data in the MCAR and MAR mechanisms using the EM and MCMC methods. The number of imputations to be performed by the MI method is determined in correlation to the efficiency table of the number of imputations that can be made for distinctive missing data rates (Schafer & Olsen, 1998).

Table 3

Relative efficiency of the number of assignments at different Missing Data Rates

Number of	Missing Data Ratio				
Assignments	%10	%30	%50	%70	%90
3	.97	.91	.86	.81	.77
5	.98	.94	.91	.88	.85
10	.99	.97	.95	.93	.92
20	1.00	.99	.98	.97	.96

When Table 3 is examined, it is seen that when the missing data is at low rates, the effect of the number of assignments is low, while the number of assignments becomes more important as the rate increases. For example, the difference between 3 assignments and 10 assignments at a rate of 10% can be interpreted as there is no need to make many assignments.

Twenty datasets were generated by imputing five data to each of the four datasets using the MI method. A total of twenty-eight datasets were created, which were imputed by three methods. Of these datasets, 20 datasets imputed by the MI were combined with DIF analysis.

Differential Item Functioning Analysis

In the final stage, DIF analysis was performed regarding the complete dataset and the native language variable. The focus group consisted of participants whose native language was not English, and the reference group consisted of native English-speaking participants. The descriptive statistics of the complete dataset are given in Table 4.

Table 4

Descriptive Statistics of the Complete Dataset

Statistics			
N_O, N_R	580	Mean	10.90
N_T	1160	Mode	12
$K_İ$	12	Median	11
$K_Ç$	5	Standard Deviation	4.27
K_T	17	Skewness	-0.021
Minimum Score	0	Kurtosis	-0.554
Maximum Score	22	Cronbach's Alpha	0.758

Notes: Focus group sample size, N_R: Reference group sample size, N_T: Total sample size, $K_İ$: Number of items dichotomous, $K_Ç$: Number of items polytomous, K_T: Total number of items

Examining the descriptive statistics showed that there were 1160 participants in total, including 580 from countries whose native language was not English in the focus group and 580 from countries whose native language was English in the reference group. There were five polytomous items (SE62095, SE62018, SE72033, SE72005, SE72920) and 12 dichotomous items (SE62099, SE62132, SE62153, SE62205, SE62190, SE72031, SE72086A, SE72086B, SE72086C, SE72086D, SE72123, SE72220) in the dataset. The mean of the booklet was 10.9, the mod was 12, the median was 11, and since they are close to each other, it can be stated that the distribution is quite normal. The skewness and kurtosis coefficients are in the range of +1, -1, which can be interpreted as not deviating excessively from the normal.

DIF analysis was performed on 28 datasets obtained by imputing values instead of missing data. The Poly-SIBTEST technique was used in the analysis. SIBTEST is a DIF determination method developed by Shealy and Stout (1993) in data scored dichotomously. Poly-SIBTEST is the general version of the SIBTEST method that can be used for dichotomous and polytomous data (Fang, 1999). Using the Poly-SIBTEST technique, DIF can be used to analyze both the item packages and the items in the data set one by one (Camilli, 2006). Since the study did not compare DIF methods, it was considered appropriate to choose a single DIF determination method. Banks and Walker (2006) used only SIBTEST in their study, obtaining good results. Sedivy et al. similarly determined DIF using the pol-SIBTEST method, stating that this was a suitable method for tests containing polytomous items. Test items are divided into two subtests, matching items and suspect items. Matching items are used as an internal matching criterion to check for group differences in target ability that is intended to be measured in DIF detection (Bolt, 2000). The analysis is carried out by dividing the test into two so that the items to be analyzed for DIF are taken into one group, and the remaining items are taken into the second group. In order to compare the performances on the items studied in the DIF, matching is done over the actual scores estimated by the total scores on the items in the second group (Gierl, 2005).

The estimate of the Poly-SIBTEST DMF index is given as

$$\hat{\beta} = \sum_{k=0}^{n_m} p_k \, (\bar{Y}_{Rk} - \bar{Y}_{Fk})$$

with k being the number of score levels in the matching items, n_m being the maximum score level in the current matching items, pk being the proportion of individuals with k matching items, Y_{Rk} being the average item score over the working group for the reference group at the current matching items, and Y_{Fk} being is the average item score for the focus group over the working group at the current matching items level.

The poly-SIBTEST statistic associated with the null hypothesis is the same as the test statistic used with the SIBTEST procedure.

$$poly - SIBTEST = \hat{\beta} \, / \hat{\sigma}(\hat{\beta})$$

$\hat{\sigma}(\hat{\beta})$ indicates that the Poly-SIBTEST DMF index is the estimated standard error.

The criteria proposed by Roussos & Stout, (1996) in interpreting the DIF effect size obtained with SIBTEST and Poly-SIBTEST are shown in Table 5. When this value is negative, the item shows DIF in favor of the focus group, and when it is positive, it shows DIF in favor of the reference group.

Table 5

β Values Interpretation Measures

DIF Level	β Value
A Level (can be ignored)	IβI <0,059
B Level (Medium Level)	0.059 ≤IβI ≤ 0,088
C Level (High Level)	IβI ≥ 0,088

DIF analyses were performed for each item one by one on 28 data sets obtained by complete data set and imputation. The reason the analysis was performed for each item separately is that the total score of the individuals in the reference and focus groups was formed by the matching items and that suspect items were not included in the matching variable. While the single item in the suspect items was analyzed, the other 16 items included in the matching items determined the total scores.
Since the number of imputations was high in the MI method, DIF analysis is given in stages.

1) With the MI method, five different imputations were made for each dataset, and 20 datasets were obtained.
2) DIF analysis was performed for the 20 datasets obtained.
3) The DIF analysis averages of the five imputations made for the missing dataset were combined.
4) The third process was repeated for all four datasets.

After DIF analysis of the data sets was obtained by imputing values instead of missing data, the results were compared with the results of the full data set DIF analysis and examined to see if there were differences in whether the items showed DIF or not and the changes in DIF levels.

Results

Findings Regarding the Complete Dataset

The poly-SIBTEST results obtained from the complete dataset to be referenced in the comparisons are presented in Table 6.

Table 6

poly-SIBTEST Analysis Findings of the Complete Dataset According to the Language Variable

Item No	β̂	p	DIF Level	Advantageous Groups	Item No	β̂	p	DIF Level	Advantageous Groups
1	-0.330	0.000*	C	Focus	10	0.169	0.000*	C	Reference
2	0.147	0.000*	C	Reference	11	-0.155	0.000*	C	Focus
3	0.212	0.000*	C	Reference	12	0.038	0.046*	A	Reference
4	0.123	0.002*	C	Reference	13	0.022	0.261		
5	-0.109	0.003*	C	Focus	14	0.008	0.671		
6	0.006	0.812			15	0.028	0.097		
7	0.042	0.133			16	0.130	0.000*	C	Reference
8	0.060	0.036*	B	Reference	17	-0.029	0.255		
9	-0.242	0.000*	C	Focus					

The complete dataset determined that one item showed DIF at A level, one item at the B level, and nine items at the C level. While items 1, 5, 9, and 11 were in favor of the focus group, the non-native English speakers, the other DIF items showed DIF in favor of the reference group, that is, native English speakers.

Findings Regarding Datasets Deleted by the Rates of 10% and 20% According to the MCAR Mechanism

According to the MCAR mechanism, the missing data obtained by deleting 10% and 20% of the data were imputed using the MCMC, EM, and MI methods. The DIF analysis of these data was performed with poly-SIBTEST, and the results are presented in Table 7.

Table 7

poly-SIBTEST Analysis Findings According to Language Variable of Datasets Generated by MCMC, EM and MI Imputation with MCAR Mechanism

Imputation Methods	MCMC				EM				MI			
Missing Data Rate	10%		20%		10%		20%		10%		20%	
Item No	β̂	p	β̂	p	β̂	p	β̂	p	β̂	p	β̂	p
1	-0.298	0.000*	-0.261	0.000*	-.291	0.000*	-0.264	0.000*	-0.285	0.000*	-0.254	0.000*
2	0.134	0.000*	0.133	0.000*	0.127	0.000*	0.122	0.000*	0.134	0.000*	0.133	0.000*
3	0.216	0.000*	0.140	0.001*	0.222	0.000*	0.150	0.000*	0.213	0.000*	0.144	0.002*
4	0.121	0.002*	0.100	0.006*	0.127	0.001*	0.097	0.006*	0.132	0.001*	0.093	0.113
5	-0.120	0.001*	-0.096	0.004*	-.128	0.000*	-0.102	0.002*	-0.130	0.002*	-0.104	0.004*
6	-0.017	0.532	0.006	0.830	-.024	0.370	-0.017	0.519	-0.010	0.906	-0.014	0.831
7	0.052	0.056	-0.005	0.842	0.054	0.046	0.013	0.614	0.042	0.489	0.060	0.246
8	0.058	0.039*	0.053	0.052	0.055	0.048*	0.046	0.091	0.048	0.034*	0.045	0.007*
9	-0.227	0.000*	-0.201	0.000*	-.223	0.000*	-0.195	0.000*	-0.223	0.000*	-0.191	0.000*
10	0.157	0.000*	0.127	0.000*	0.173	0.000*	0.144	0.000*	0.163	0.000*	0.126	0.000*
11	-0.137	0.000*	-0.133	0.000*	-.147	0.000*	-0.117	0.000*	-0.148	0.000*	-0.113	0.000*
12	0.027	0.145	0.021	0.246	0.025	0.183	0.026	0.155	0.031	0.047*	0.022	0.060*
13	0.024	0.202	0.015	0.443	0.022	0.245	0.014	0.454	0.020	0.247	0.016	0.306
14	0.017	0.380	0.000	0.988	0.009	0.614	0.000	0.988	0.014	0.376	-0.007	0.919
15	0.028	0.110	0.024	0.153	0.024	0.164	0.026	0.108	0.030	0.036*	0.020	0.271
16	0.122	0.000*	0.119	0.000*	0.111	0.000*	0.139	0.000*	0.116	0.000*	0.120	0.000*
17	-0.045	0.071	0.005	0.845	-.038	0.132	0.007	0.779	-0.046	0.068	-0.014	0.409

According to the poly-SIBTEST results of the data set generated by imputing MCMC to the 10% MCAR mechanism data set, it was observed that item 12 of the items showing DIF in the complete data set did not show DIF, and item 8 was a different DIF level. When compared with the results of the complete data set, items 8 and 12 of the items with DIF did not show DIF. In both of the data sets with missing data in both ratios and imputed with the MCMC method, none of the items without DIF in the complete data set showed DIF. Of the items with DIF in the complete data set, 91% and 82% showed DIF at 10% and 20%, respectively.

According to the poly-SIBTEST results of the data set generated with the EM imputation, the data set with the 10% MCAR mechanism showed that the 12th item among the items with DIF in the full data set did not show DIF, while the seventh item without DIF showed DIF, unlike the results of the full data set. In addition, the DIF level of item 8 showed a difference. The poly-SIBTEST results of the data set created by imputing EM to the 20% MCAR mechanism data set determined that nine items showed DIF at the C level, and unlike the results of the complete data set, the eighth and 12th of the items with DIF did not show DIF. In the complete data set of data sets, which contained 10% and 20% missing data and was imputed with EM method, 83% and 100% of the items without DIF did not show DIF. 91% and 82% of the items with DIF in the complete data set showed DIF, respectively, in the rates of 10% and 20%. According to the poly-SIBTEST results of the data set created with the MI imputation, it was determined that the 15th item, which did not show DIF in the full data set, showed DIF at the A level, and the eighth item had a different DIF level, unlike the full data set in the 10% MCAR mechanism data

set. In the data set generated by imputing MI to the rate of 20% MCAR mechanism data set, unlike the results of the complete data set, the fourth item with DIF did not show DIF, the 15th item without DIF showed DIF at the A level, and the level of DIF for the eighth item showed a difference. In both data sets with 10% and 20% missing data and imputed by the MI method, 83% of the items without DIF in the complete data set did not show DIF. Of the items with DIF in the complete data set, 100% and 91% showed DIF, respectively, at the rates of 10% and 20%.

Findings Regarding Datasets Deleted by the Rate of 10% and 20%

According to the MAR mechanism, data were imputed to missing data obtained by deleting 10% and 20% of them using the MCMC, EM, and MI methods. The DIF analysis of these data were performed with poly-SIBTEST, and the results are shown in Table 8.

Table 8

poly-SIBTEST Analysis Findings According to Language Variable of Datasets Generated by MCMC, EM, and MI Imputation with MAR Mechanism

Imputation Methods	MCMC				EM				MI			
Missing Data Rate	10%		20%		10%		20%		10%		20%	
Item	β	p	β	p	β	p	β	p	β	p	β	p
1	-0.298	0.000*	-0.291	0.000*	-0.288	0.000*	-0.278	0.000*	-0.304	0.000*	-0.304	0.000*
2	0.154	0.000*	0.136	0.000*	0.148	0.000*	0.159	0.000*	0.135	0.001*	0.110	0.004*
3	0.188	0.000*	0.177	0.000*	0.193	0.000*	0.177	0.000*	0.171	0.000*	0.191	0.000*
4	0.114	0.003*	0.118	0.001*	0.124	0.002*	0.131	0.000*	0.127	0.005*	0.085	0.028*
5	-0.093	0.008*	-0.120	0.000*	-0.107	0.002*	-0.148	0.000*	-0.088	0.015*	-0.104	0.010*
6	0.008	0.768	-0.018	0.501	0.002	0.944	0.018	0.499	0.013	0.871	0.024	0.607
7	0.012	0.668	0.046	0.089	0.017	0.531	0.017	0.519	0.027	0.189	0.053	0.030*
8	0.041	0.137	0.056	0.038*	0.036	0.185	0.056	0.035*	0.046	0.174	0.067	0.035*
9	-0.200	0.000*	-0.172	0.000*	-0.199	0.000*	-0.175	0.000*	-0.206	0.000*	-0.201	0.000*
10	0.151	0.000*	0.104	0.000*	0.148	0.000*	0.106	0.000*	0.150	0.000*	0.141	0.000*
11	-0.142	0.000*	-0.106	0.000*	-0.121	0.000*	-0.111	0.000*	-0.133	0.000*	-0.146	0.000*
12	0.028	0.145	0.001	0.941	0.017	0.390	-0.004	0.841	0.028	0.131	0.025	0.229
13	0.004	0.838	-0.008	0.687	-0.004	0.825	-0.013	0.484	0.011	0.326	0.011	0.341
14	-0.002	0.896	-0.014	0.460	-0.009	0.606	-0.018	0.333	0.006	0.693	0.008	0.702
15	0.023	0.180	0.012	0.497	0.014	0.412	0.002	0.887	0.029	0.112	0.010	0.819
16	0.090	0.001*	0.092	0.001*	0.095	0.000*	0.099	0.000*	0.101	0.000*	0.107	0.000*
17	-0.006	0.814	0.043	0.069	-0.005	0.851	0.072	0.002*	-0.033	0.092	-0.013	0.894

According to the poly-SIBTEST results of the data obtained by imputing MCMC to the data with 10% MAR, it was observed that the eighth and 12th items with DIF in the complete data set did not show DIF. When the results of the data set generated by imputing MCMC to the data set with a 20% MAR mechanism in the complete data set were examined, it was seen that the 12th item with DIF did not show DIF and the eighth item had a distinctive DIF level. In both of the data sets with 10% and 20% missing data and imputed with the MCMC method, none of the items without DIF in the complete data set showed DIF. Of the items with DIF in the complete data set, 82% and 91% of them showed DIF, respectively, at the rates of 10% and 20%.

According to the results of the poly-SIBTEST of the data set generated by the EM imputation, compared with the results of the complete data set, it was seen that the eighth and 12th items with DIF did not show DIF. It was determined that the 12th item with DIF did not show DIF, and the DIF levels of the eighth and 17th items showed a difference in the data set created by imputing EM to the 20% MAR

mechanism data set, unlike the results of the complete data set. In both data sets with 10% and 20% missing data and data sets that imputed with the EM method in order, all and 83% of the items without DIF in the complete data set did not show DIF. In both data sets, 82% of the items with DIF in the complete data set showed DIF.

According to the poly-SIBTEST results of the data set generated with the imputation of MI, in the 10% MAR mechanism data set, unlike the complete data set, the eighth and 12th items with DIF did not show DIF. In the data set generated by imputing MI to the data set with a 20% MAR mechanism, it could be seen that the 12th item with DIF did not show DIF, the seventh item without DIF showed DIF, and the fourth item had different DIF levels. While 83% of the items without DIF in the complete data set did not show DIF in both of the data sets consisting of 10% and 20% missing data and imputed with the MI method, the items with DIF in the complete data set were at the rates of 10% and 20%, respectively, and 100% and 91% of them showed DIF again. In the complete data set of the data sets with 10% and 20% missing data and imputed with the MI method in order; all and 83% of the items without DIF did not show DIF. 82% and 91% of the items with DIF in the complete data set showed DIF.

Findings Regarding the Distribution of Items Displaying and Not Displaying DIF

The distributions of the 12 items displaying DIF in the complete dataset as a result of imputing values using the EM, MCMC, and MI methods are presented in Table 9.

Table 9

Distributions of the Items Displaying DIF in the Complete Dataset According to the Missing Data Mechanisms and Missing Data Imputation Methods in the Missing Data

Missing Data Mechanism		MCAR						MAR					
Missing Data Imputation Method		EM		MCMC		MI		EM		MCMC		MI	
Missing Data Rate	DIF Level	f	%	f	%	f	%	f	%	f	%	f	%
10%	A	0	0	0	0	1	100	0	0	0	0	0	0
	B	1	100	1	100	1	100	0	0	0	0	0	0
	C	9	100	9	100	9	100	9	100	9	100	9	100
20%	A	0	0	0	0	1	100	0	0	0	0	0	0
	B	0	0	0	0	1	100	1	100	1	100	1	100
	C	9	100	9	100	8	89	9	100	9	100	9	100

A and B levels of DIF could only be determined by imputation with the MI method at a missing data rate of 20% under the MCAR mechanism. While the B level can be determined in all methods at the rate of 10% missing data under the MCAR mechanism, the A level could only be determined by the MI method. While under the MAR mechanism, items displaying DIF at levels A and B at a rate of 10% missing data could not be determined in all three methods, and only the A level could not be determined at a rate of 20%. In general, as the rate of missing data increased, the inability to correctly identify items with DIF increased. Similarly, Tamcı (2018) stated that the status of DIF items as a result of imputation with MI and EM methods shows good results in some circumstances while the status of items with DIF as a result of imputation was badly affected in an increase in the missing data rate. The items displaying DIF at the B and C levels had similar outcomes to the complete dataset in all methods at a missing data rate of 20% under the MAR mechanism and a missing data rate of 10% under the MCAR mechanism.

As for the C level DIF, the same results were obtained with the complete dataset in all other conditions except for the imputation by the MI method at a missing data rate of 20% under the MCAR mechanism. Under the MAR mechanism, all C-level DIF items were identified by all methods. When we look at the situation of whether the items displaying DIF still displayed DIF as a results of value imputation, it was found that 10% of the data were missing under the MCAR mechanism, and DIF was identified in all of

the items displaying DIF in the complete dataset at all DIF levels when the imputation was made with the MI method.

The distributions of the six items that do not display DIF in the complete dataset as a result of value imputation are given in Table 10.

Table 10

Distributions of the Items Not Displaying DIF in the Complete Dataset According to the Missing Data Mechanisms and Missing Data Imputation Methods in the Missing Data

Missing Data Mechanism	MCAR						MAR					
Missing Data Imputation Method	EM		MCMC		MI		EM		MCMC		MI	
Missing Data Rate	f	%	f	%	f	%	f	%	f	%	f	%
10%	5	83,3	6	100	5	83,3	6	100	6	100	6	100
20%	6	100	6	100	5	83,3	5	83,3	6	100	5	83,3

In all circumstances, the MCMC method correctly identified all items in the complete dataset that were not displaying DIF. DIF was observed in one item in which no DIF was observed in the complete dataset when imputing it using the EM and MI methods at a missing data rate of 10% under the MCAR mechanism and at a missing data rate of 20% under the MAR mechanism.
When examining whether items without DIF display DIF, the MCMC method was usually found to be preferable to the other methods in all conditions. Garrett (2009) stated that under the MCAR mechanism, the fact that items without DIF did not show DIF as a result of imputation with the MI method was better than the other methods used in their study.

Discussion and Conclusion

This study examined how DIF results differentiate according to the DIF level, missing data rate, and missing data mechanism when data imputation is performed using the MCMC, EM, and MI methods and considering the native language variable. The findings obtained from this study were presented by examining how the results obtained from the complete dataset and the MCMC, EM, and MI methods change in each missing data mechanism and missing data rate.

According to the missing data rate condition discussed in the research, it was seen that when the MCMC method was used, the increase in the missing data rate under the MCAR mechanism decreased in the correct identification of DIF, and the increase in the missing data rate under the MAR mechanism showed improvement in detecting DIF correctly. If there is an increase in the rate of missing data under the MAR mechanism, the use of the MCMC method was found to be more suitable. Finch (2011b) stated in her study that assigned stochastic regression imputation, logistic regression, and MI with zero imputation; for MI, the error decreased as the percentage of missing data increased while there was no decrease or increase in other methods. Nichols et al. (2022) stated that when the missing data rate is larger than 10% much larger magnitudes of DIF estimation error were observed.

Considering DIF levels, the MCMC and EM methods had good results in items with DIF at the C level under both the MCAR and the MAR mechanisms. While the MI method performed well with a 10% missing data rate, incorrect identifications were observed at the C level when the missing data rate increased. Based on these results, while all three methods can be preferred at a 10% missing data rate at a high DIF level (C), the EM and MCMC methods were found to be preferable to the MI method with a 20% missing data rate. The EM and MCMC methods with a missing rate of 20% under the MCAR mechanism and three methods with a missing rate of 10% under the MAR mechanism had the same results. In the MCMC and EM methods, it was observed that items with A and B levels DIF could not

be identified in the 20% MCAR and 10% MAR mechanisms, and while the A level DIF could not be identified in the 10% MCAR and 20% MAR mechanisms, it identified the B level as DIF at a lower level. While the MI method was some successful in identifying the A level in all missing data rates under the MCAR mechanism, it was determined the B level was identified at a low level, like other methods. The 10% MAR mechanism could determine DIF only at the C level and was unable to determine DIF at other levels. The finding was that the MI method is preferable to other methods for identifying A-level DIF. A level DIF was seen in items without DIF with the MI method in the MCAR mechanisms and 20% MAR mechanism.

With the increase in the rate of missing data in the MCAR mechanism, incorrect determinations were observed in the determination of substances with DIF. As the missing data rate increased, the EM and MCMC methods were found to be preferable over the MI method. Items with DIF were best identified under the MCAR mechanism when imputing them with the MI method at a missing data rate of 10%. While all items with DIF were determined by this method, the level of only one item with DIF was determined distinctively. Similarly, Finch (2011a) mentioned that the results obtained when he imputed using the MI method under the MCAR mechanism were compatible with the complete data, and, in his other study, Finch (2011b) stated that the results obtained when he imputed using the MI method again, except for the 10% missing data condition, were also compatible with the complete data. In the other study, Finch stated that when the type I error rates in the MAR mechanism were examined, the error rates for MI were lower than for the other two methods (zero imputation and stochastic regression imputation). On the other hand, Garrett (2009), in his study investigating the effects of MI and mean value imputing methods and the MH and ordinal LR methods, which are DIF determination methods suggested the use of MI, one of the methods of coping with missing data, when both DIF determination methods are used.

When the situations of whether items without DIF displayed DIF were examined, DIF was detected in items without DIF when using MI method at the rate of 10% and 20% missing data and EM methods at the rate of 10% missing data under the MCAR mechanism. Under the MAR mechanism, DIF was identified in items without DIF when EM and MI methods were used at a missing data rate of 20%. The MCMC method showed good results by identifying all items without DIF at both missing data rates under the MCAR mechanism, the EM method at a missing data rate of 20%, the MCMC method at both missing data rates under the MAR mechanism, and the EM and MI methods at a missing data rate of 10% without DIF. These indicators support the results of the study conducted by Tamcı (2019), which indicated that the MI method works better on items that do not display DIF than the EM method, and all items that did not display DIF for all other conditions at the missing data rate of 30% came out without DIF in the MI method.

In light of these results, it cannot be said that only one method is good. Different results were obtained for various conditions. For example, Finch (2011b) stated that study results suggested that the relationships between the different factors manipulated were complex with no one method emerging as fixed in all cases; however, listwise deletion consistently produced results similar to those obtained with the complete data set across simulated conditions. When the missing data rate exceeds the 10% threshold, Nichols et al. (2022) recommended MICE in their study due to missing data when testing for DIF. However, they stated that the methods were unable to completely eliminate the observed error due to missing data. Therefore, whichever method is used by the researchers, they should interpret the results carefully.

Recommendations Based on Research Results

Since the results of the MCMC and EM methods are found to be more similar to the complete dataset in cases where there are items displaying C-level DIF, these methods are recommended for imputing missing data in DIF studies. The MI method can be preferred when there are items with DIF at A and B levels.

In cases where the rate of missing data is high in the correct detection of items without DIF, it is recommended that the MCMC and EM methods be used under the MCAR mechanism and the MCMC method under the MAR mechanism.

For low missing data rates, it is recommended that the MCMC method be used under the MCAR mechanism and all three methods under the MAR mechanism.

In the identification of items with DIF, it is recommended that the MI method be used in cases where the missing data rate is high under MCAR and MAR mechanisms, and all three methods should be used with low missing data rates.

Recommendations Based on Subsequent Research

Based on the results of the research, in cases where there are items showing C-level DIF, the results of MCMC and EM methods are found to be more similar to the complete data set, so it is recommended that these methods be chosen for imputing missing data in DIF studies. When there are items with DIF at A and B levels, the MI method can be preferred.

In cases where the rate of missing data is high in the correct determination items without DIF, it is recommended that MCMC and EM methods be used under the MCAR mechanism and the MCMC method under the MAR mechanism.

For low missing data rates, it may be recommended that the MCMC method be used under the MCAR mechanism and all three methods under the MAR mechanism. In the detection of items with DIF, it can be recommended that the MI method be used in cases where the missing data rate is high under MCAR and MAR mechanisms, and all three methods should be used with low missing data rates.

In this study, a single DIF method was used. Despite DIF determination methods having similar results in general, as Gök et al. (2010) stated, there is no complete harmony between the methods, and it is recommended that different DIF methods be used, such as LR, the IRT probability rate, which has used different matching criteria, algorithms, and breakpoints.

In this study, the sample size was fixed. The mechanisms for missing data and methods of dealing with missing data at different sample sizes could be studied. Of the methods for dealing with missing data, three imputation methods were used from the class of probabilistic and translational data imputation. The number of methods can be increased, and comparisons can be made by using methods based on the deletion and simple imputation. DIF analysis was conducted using a CTT-based method. DIF analysis can also be conducted with IRT-based methods and techniques. In addition, the study can be expanded by increasing conditions for the test length, uniform and non-uniform DIF, focus-reference group rates, missing data rate, and methods for dealing with missing data.

Declarations

Author(s) contribution: Leyla Burcu DİNÇSOY-Investigation, methodology, visualization, software, formal analysis, and writing-original draft. Hülya KELECİOĞLU-Developing process, methodology, resources, supervision and validation.

Conflict of Interest: No potential conflict of interest was reported by the authors.

Ethical Approval: This research study complies with research publishing ethics. Secondary data were used in this study. Therefore, ethical approval is not required.

References

Allison, P. D. (2002). *Missing data*. Sage.
Alpar, R. (2021). *Çok değişkenli istatistiksel yöntemler.* Detay.

Banks, K. (2015). An introduction to missing data in the context of differential item functioning. *Practical Assessment, Research & Evaluation*, *20*(12), 12. https://doi.org/10.7275/FPG0-5079

Banks, K., & Walker, C. M. (2006). *Performance of SIBTEST when focal group examinees have missing data*. Paper presented at the annual meeting of the National Council on Measurement in Education.

Bolt, D. M. (2000). A SIBTEST approach to testing DIF hypotheses using experimentally designed test items. *Journal of Educational Measurement, 37*(4), 307-327. https://doi.org/10.1111/j.1745-3984.2000.tb01089.x

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.). American Council on Education & Praeger Publishers.

Çüm, S., Demir, E. K., Gelbal, S., & Kışla, T. (2018). A comparison of advanced methods used for missing data imputation under different conditions. *Mehmet Akif Ersoy University Journal of Education Faculty, 45*, 230-249. https://doi.org/10.21764/maeuefd.332605

Demir, E. (2013). Item and test parameters estimations for multiple choice tests in the presence of missing data: The case of SBS. *Journal of Educational Sciences Research, 3*(2), 47-68. http://dx.doi.org/10.12973/jesr.2013.324a

Emenogu, B. C., Falenchuck, O., & Childs, R. A. (2010). The effect of missing data treatment on Mantel-Haenszel DIF detection. *The Alberta Journal of Educational Research, 56*(4), 459-469. https://doi.org/10.11575/ajer.v56i4.55429

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

Falenchuk, O., & Herbert, M. (2009). Investigation of differential non-response as a factor affecting the results of Mantel-Haenszel DIF detection. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Fang, T. (1999). *Detecting DIF in polytomous item responses* [Doctoral dissertation, University of Ottawa]. https://ruor.uottawa.ca/handle/10393/8495?locale=fr

Finch, H. (2011a). 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. https://doi.org/10.1080/08957347.2011.607054

Finch, H. W. (2011b). The impact of missing data on the detection of nonuniform differential item functioning. *Educational and Psychological Measurement, 71*(4), 663-683. https://doi.org/10.1177/0013164410385226

Garrett, P. (2009). *A Monte Carlo study investigating missing data, differential item functioning and effect size* [Doctoral thesis, Georgia State University]. https://doi.org/10.57709/1060078

Gierl, M. J. (2005). Using dimensionality-based DIF analysis to identify and interpret constructs that elicit group differences. *Educational Measurement: Issues and Practice, 24*, 3-14. https://doi.org/10.1111/j.1745-3992.2005. 00002.x

Gill, J. (2002). *Bayesian methods: A social and behavioral sciences approach*. Chapman and Hall.

Gök, B., Kelecioğlu, H., & Dogan, N. (2010). The comparison of Mantel-Haenszel and logistic regression techniques in determining the differential item functioning. *Education and Science, 35*(156), 3-16.

Hambleton, R. K., Clauser, B. E., Mazor, K. M. & Jones, R. W. (1993). *Advances in the detection of differentially functioning test items*. University of Massachusetts, School of Education. http://files.eric.ed.gov/fulltext/ED356264.pdf

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the DIF analysis. *Test Validity*, 129-145.

Karasar, N. (2011*). Bilimsel araştırma yöntemleri*. Nobel.

Little, R., & Rubin, D. (2020). *Statistical analysis with missing data* (4th ed.). Wiley.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*, 297-334. https://doi.org/10.1177/014662169301700401

Nichols, E., Deal, J. A., Swenor, B. K., Abraham, A. G., Armstrong, N. M., Bandeen-Roche, K., Carlson, M.C., Grisworld, M., Lin, F. R., Mosley, T. H., Ramulu, P. Y., Reed, N. S., Sharrett, A. R., & Gross, A. L. (2022). The effect of missing data and imputation on the detection of bias in cognitive testing using differential item functioning methods. *BMC Medical Research Methodology, 22*(1), 1-12. https://doi.org/10.1186/s12874-022-01572-2

Özgüven, İ. E. (2017). *Psikolojik testler*. Nobel.

Robitzsch, A., & Rupp, A. A. (2009). Impact of missing data on the detection of differential item functioning: The case of Mantel-Haenszel and Logistic Regression analysis. *Educational and Psychological Measurement, 69*(1), 18-34. https://doi.org/10.1177/0013164408318756

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error Performance. *Journal of Educational Measurement, 33*(2), 215-230. https://doi.org/10.1111/j.1745-3984.1996.tb00490.x

Rousseau, M., Bertrand, R., & Boiteau, N. (2006). *Impact of missing data treatment on the efficiency of DIF methods*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Rubin D. B. (1976). *Inference and missing data*. *Biometrika*, *72*, 359-364.

Selvi, H., & Alıcı, D. (2018). Investigating the impact of missing data handling methods on the detection of differential item functioning. *International Journal of Assessment Tools in Education, 5*(1), 1-14. https://doi.org/10.21449/ijate.330885

Sedivy, S. K., Zhang, B., & Traxel, N. M. (2006). *Detection of differential item functioning with polytomous items in the presence of missing data*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*, *33*(4), 545-571.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147-177. https://doi.org/10.1037/1082-989X.7.2.147

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*(2), 159-194. https://doi.org/10.1007/BF02294572

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Tabachnick, B., & Fidell, L. (1996). *Using multivariate statistics* (3rd ed.). Herper Collins College Publishers.

Tamcı, P. (2018). *Kayıp veriyle başa çıkma yöntemlerinin değişen madde fonksiyonu üzerindeki etkisinin incelenmesi [Investigation of the impact of techniques of handling missing data on differential item functioning]* (Thesis No. 517260) [Master's thesis, Hacettepe University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Toka, O. (2012). *Kayıp veri durumunda sağlam kestirim [Robust estimation in case of missing data]* (Thesis No. 321449) [Master's thesis, Hacettepe University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Turgut, M. F., & Baykul, Y. (2012). *Eğitimde ölçme ve değerlendirme []*. Pegem.

Van Buuren, S. (2012). *Flebitler imputation of missing data.* CRC Press.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Directorate of Human Resources Research and Evaluation, Department of National Defense.