ISSN: 1309 – 6575
**Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi**
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(1); 95-105

# The Impact of Missing Data on the Performances of DIF Detection Methods

Rabia AKCAN*     Kübra ATALAY KABASAKAL**

**Abstract**

This study analyzed the impact of missing data techniques on performances of two differential item functioning (DIF) detection methods (Mantel Haenszel and Multiple Indicator and Multiple Causes) under missing completely at random missing data mechanism. Percentage of missing data was set at 5% and 15%. Zero imputation, listwise deletion and fractional hot-deck imputation were used to handle missing data. The data set of the study consisted of 17 items in the S12 item cluster of Programme for International Student Assessment (PISA) 2015 science test. Results showed that fractional hot-deck imputation produced the best results in identifying DIF items in all conditions and it had also the closest DIF values to the values obtained from complete data set. It was also found that multiple indicator and multiple causes method was more adversely affected than Mantel Haenszel by the presence of missing data.

*Keywords: Differential item functioning, Mantel Haenszel, MIMIC, missing data.*

## Introduction

Missing data is a frequently encountered problem in quantitative research studies. Since standard statistical methods were designed for complete data sets, missing values create a significant problem for the researchers. Generally, researchers use various ad hoc methods to handle missing data before the analysis. An example of these strategies is discarding the cases with missing data (i.e., listwise deletion). Replacing missing values with variable mean is another method. Yet, these traditional methods can lead to significant bias in sample statistics (Peugh & Enders, 2004).

The rate of missing data, missing data mechanism and patterns of missing data should be considered in order to decide on the method to handle missing data. Rate of missing data is directly associated with the quality of statistical inferences. There is not a specified criterion in the literature with respect to a reasonable missing data rate to get valid statistical inferences (Dong & Peng, 2013). However, it is seen that the rate of missing data has mostly varied between 0% and 30% in previous studies (Banks & Walker, 2006; Finch, 2011a; Finch, 2011b; Robitzsch & Rupp, 2009; Rousseau et al., 2004).

As previously stated, another aspect of handling missing data is to take the missing data mechanism into account. Rubin (1976) classified missing data mechanisms into three types: Missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Within the context of item responses, MCAR indicates that some examinees leave the item blank in a completely random way without a systematic mechanism related to the missingness. Data are MAR when the probability of an observation which includes missing data is directly connected with a measurable variable. The fact that male students' probability of leaving an item blank is higher than female students would be an illustration of MAR mechanism. MNAR mechanism refers to the case in which probability of being missing is related to the value of the variable itself. In this case, an examinee might leave the item unanswered as they do not know the answer (Finch, 2011b).

* Teacher, Ministry of National Education, Afyonkarahisar-Türkiye, eltrabia42@hotmail.com, ORCID ID: 0000-0003-3025-774X
** Assoc. Prof., Hacettepe University, Faculty of Education, Ankara-Türkiye, kkatalay@gmail.com, ORCIDID: 0000-0002-3580-5568

To cite this article:
Akcan, R., & Atalay-Kabasakal, K. (2023). The impact of missing data on the performances of DIF detection methods. *Journal of Measurement and Evaluation in Education and Psychology, 14*(1), 95-105. https://doi.org/10.21031/epod.1183617

*Received: 3.10.2022*
*Accepted: 22.03.2023*

Missing data can affect quantitative research severely and may cause bias in parameter estimates, reduced statistical power, inflated standard errors and information loss (Dong & Peng, 2013). Therefore, it is essential for the researchers to investigate the impact of missing data on statistical techniques. Of particular concern in this research is the effect of missing data on the detection of differential item functioning (DIF), which causes systematic errors and reduces validity, with different methods. What follows is a brief overview of DIF and the methods used in this study.

DIF has received considerable attention as a result of the increased reliance on standardized achievement testing for evaluating the progress in education.  The need to provide accurate assessment for all the examinees comes with great responsibility for psychometricians. Items intended to measure reading skills, for instance, must be suitable for the use with the students from various groups (e.g. gender, ethnicity etc.) to get meaningful score interpretations (Finch & French, 2007). If an item functions differently in a focal group compared with a reference group after controlling for differences in levels of performance on a latent trait (e.g., ability) of interest, it means the item shows DIF (Holland & Wainer, 1993; Scheuneman, 1979).DIF can be categorized into two broad types: Uniform and nonuniform. Uniform DIF is present when one of two groups has uniformly greater probability of answering an item correctly across all ability levels (Finch, 2005). Nonuniform DIF occurs when members of one group have greater probability in responding to an item correctly for some levels of the ability being measured, while they have lower probability for the other levels of the ability (Camilli & Shepard, 1994).

DIF detection methods can broadly be examined under two headings: (1) Classical Test Theory (CTT) and Item Response Theory (IRT). However, Camilli and Shepard (1994) highlighted that Confirmatory Factor Analysis (CFA) methods can be used to identify DIF as well. Previous studies in the field of DIF in the presence of missing data have mostly focused on CTT and IRT methods rather than CFA (Banks & Walker, 2006; Finch, 2011a; Finch, 2011b; Robitzsch & Rupp, 2009; Rousseau et al., 2004). In this respect, we decided to use Mantel Haenzsel, a widely accepted method in literature based on CTT, and multiple indicator and multiple causes which is a CFA method becoming popular recently.

## MIMIC

Multiple indicator and multiple causes (MIMIC) method is based on CFA and has received growing attention on DIF detection. The fundamental technique underlying DIF assessment with MIMIC models includes estimation of both direct and indirect effects for a grouping variable. The indirect effect shows whether there is a difference in the mean of latent trait across the groups, thereby explains the group differences on the latent trait. The direct effect shows whether response probabilities differ across the groups. In the DIF framework, MIMIC model can be written as (Finch, 2005):

$$y_i^* = \lambda_i \eta + \beta_i z_k + \varepsilon_i, \qquad (1)$$

where

$y_i^*$= latent response variable;

$\lambda_i$ = factor loading for variable i;

$\eta$ = latent trait;

$\beta_i$ = slope relating the group variable with the response;

$\varepsilon_i$ = random error; and

$z_k$ = a dummy variable showing group membership.

Previous simulation studies investigating DIF with MIMIC method have shown that under most circumstances, MIMIC method performed as efficiently as or better than the other methods (SIBTEST, MH, LR etc.) with regard to type I error rate and power (e.g., Finch, 2005; Uğurlu & Atar, 2020; Woods, 2009). Missing data is a significant factor in the performances of statistical methods. Therefore, the impact of missing data on the DIF detection with MIMIC model is an important issue to be considered.

## Mantel Haenszel

Mantel Haenszel (MH) statistic, proposed by Holland and Thayer (1988), might be the most commonly used among contingency table methods. With this method, probability of success on the item is compared for the members of two groups that are matched on the ability being measured. Firstly, respondents are divided into levels depending on the ability. Total test score is generally used for matching the respondents. For each score level, a 2x2 table is then created as in Table 1 (Clauser & Mazor, 1998).

**Table 1**

*Data Organization in MH Method*

| Group | 1 =Correct | 0=Incorrect | Total |
|---|---|---|---|
| Reference | $A_j$ | $B_j$ | $N_{Rj}$ |
| Focal | $C_j$ | $D_j$ | $N_{Fj}$ |
| Total | $M_{1j}$ | $M_{0j}$ | $T_j$ |

MH statistic gives odds ratio (α), the ratio of the odds that reference group will respond to the studied item correctly to those for the focal group (Clauser & Mazor, 1998). Odds ratio is given in the equation (2).

$$\alpha = \frac{\sum_j A_j D_j / T_j}{\sum_j B_j C_j / T_j} \qquad (2)$$

Holland & Thayer (1988) recommended a logistic transformation to make interpretation of odds ratio easier. First, log of α is taken in order that the scale is symmetric around zero. Then, resulting value is multiplied by – 2.35 which produces $\Delta_{MH}$ (Clauser & Mazor, 1998).Zieky (1993) classified $\Delta_{MH}$ statistic into three categories: $|\Delta_{MH}| < 1$ shows negligible DIF (A level), $1 \leq |\Delta_{MH}| \leq 1.5$ shows moderate DIF (B level) and $|\Delta_{MH}| \geq 1.5$ shows large DIF (C level).

Returning briefly to missing data, it is obvious that presence of missing data is an important issue with regard to the DIF detection. However, commonly used DIF detection methods such as MH, SIBTEST and Logistic Regression (LR) are not capable of handling missing data. Hence, missing data handling methods used for the analysis might cause bias. Choice of missing data method may create DIF when there is no DIF in the item or eliminate DIF when it is actually present (Banks, 2015). When the choice of missing data handling method is inconvenient, erroneous decisions can be made based on DIF results which may prevent meaningful test score interpretations.

Researchers have attempted to assess the impact of missing data on DIF detection via simulation studies (Banks & Walker, 2006; Finch, 2011a; Finch, 2011b; Garrett, 2009; Robitzsch & Rupp, 2009) or studies with real data (Rousseau et al., 2004; Tamcı, 2018). Most of these studies have focused on the widely used DIF detection methods such as SIBTEST, MH or LR. Emenogu et al. (2010) used both real and simulated data to investigate the impact of zero imputation (ZI), listwise deletion (LD) and analysis wise deletion on MH method. They reported that ZI produced false DIF regardless of the matching criterion used in the study and LD led to a significant decrease in sample size and the power of MH method.

Finch (2011b) also included IRT-LR in his study along with crossing SIBTEST and LR. This study has assessed the efficacy of ZI, LD, multiple imputation (MI) and stochastic regression imputation (SRI) on DIF detection. LD was recommended as a traditional missing data handling method for each DIF method and MI was the imputation method recommended in the study.

In recent years, there has been a growing amount of literature on the DIF detection with MIMIC, a CFA-based DIF detection method (Finch, 2005; Jin & Chen, 2020; Montoya & Jeon, 2020; Shih & Wang, 2009; Uğurlu & Atar, 2020; Woods, 2009). Missing data can affect any type of analysis including CFA (Harrington, 2009). Therefore, this study uses MIMIC method along with MH which is a broadly accepted method in the literature.

Zero imputation, listwise deletion and fractional hot-deck imputation (FHDI) were chosen as missing data handling method in the current study because the first two were widely used in prior research and far too little attention was paid to the last one. For ZI, all missing responses were replaced with 0. For LD, all individuals who had incomplete data responses were deleted. In FHDI, proposed by Kalton and Kish (1984) and investigated by Kim and Fuller (2004), M imputed values are created for each missing value, however, after fractional imputation a single data set is obtained as the output. Fractional weights are assigned to imputed values. The purpose of FHDI is to perform hot deck imputation efficiently (Im et al., 2015). FHDI was extended by Im et al. (2015) in two ways. First, in this new version of FHDI imputation cells are not required to be made in advance. Second, the proposed FHDI method is applied multivariate missing data with arbitrary missing patterns. In this paper, we used extension of FHDI proposed by Im et al. (2015) which is available in R software.

## Purpose of the Study

DIF detection is an increasingly important area in test development and validity of standardized achievement tests which contribute to the development of educational policies (Zumbo, 2007). PISA (The Program for International Student Assessment), which enables comparison of students' achievement from different countries and languages and directs educational policies of these countries, is one of the important international standardized tests. Missing data can also be a problem in PISA application as with many other tests (e.g., Emenogu et al., 2010; Tamcı, 2018).

As already stated, traditional DIF detection methods cannot handle missing data. However, it is natural to have missing data in many educational or psychological tests. In this case, solving the missing data problem before DIF analysis becomes essential. Several studies investigating the missing data and DIF detection demonstrated that choice of missing data treatment method or type of missing data can have an influence on the DIF detection methods' performances (Finch, 2011a; Robitzsch & Rupp, 2009). This study therefore set out to assess the performances of DIF detection methods in PISA application in the presence of missing data. The leading research question in this investigation was as follows: What is the impact of (a) different missing data handling methods under (b) MCAR missing data mechanism and (c) different missing data percentages on the performances of the MH and MIMIC DIF detection methods?

## Methods

This study aims to determine the impact of three missing data techniques on the performances of DIF detection methods under MCAR missing data mechanism. In this respect, this study is a descriptive study as it describes the existing situation as precisely as possible (Fraenkel et al., 2012).

## Data Set

The data set consists of 17 items in the S12 item cluster of PISA 2015 science test. 1099 students from Finland who responded to all the items in the test were recruited as the sample of the study. Gender DIF studies are commonly carried out in international tests. However, gender DIF was not studied to make inferences on gender in this study. Different size of focal and reference groups might be another variable and affect the performances of missing data handling methods. As a result of this, Finland data set (1362 students) was chosen as the sample in the present study because the number of reference (female) and focal (male) groups was almost equal after discarding missing data.

## Data Analysis

Data set includes 16 binary scored items and a partially scored item (CS637Q02S). This item was coded as 1-0 (full and partial point coded were as 1 and others were coded as 0) by the researchers and analyses were carried out on 17 items. A complete data set of 1,099 people (550 female and 549 male students) was obtained by discarding the missing data from the data set. After gender-based DIF analyses on complete data set were conducted with MH and MIMIC methods, results were recorded to be used as reference. Following DIF analyses, missing responses were created on complete data set by deleting data under MCAR. Missing responses under MCAR mechanism were created by selecting responses randomly from all items and all responses (0-1) for both groups. As the percentage of missing data mostly ranged between 0% and 30% in prior research (Banks & Walker, 2006; Finch, 2011a; Finch, 2011b; Robitzsch & Rupp, 2009; Tamcı, 2018), the percentage of missing data in the current research was set at 5% and 15%. Missing data were then dealt with ZI, LD and FHDI methods. DIF analyses were performed on these data sets. Finally, a comparison was made between reference DIF results and the results obtained from data sets that were completed with missing data handling methods. Whether numbers, levels or directions of DIF items in complete data set have changed or not was investigated. Pearson correlations of MH and MIMIC DIF statistics in all conditions were also examined. “MplusAutomation” (Hallquist & Wiley, 2018) and “difR” (Magis et al., 2010) packages were used for DIF analyses with MIMIC and MH methods respectively. Missing responses were generated in R through adapting the missing data codes written by Doğanay Erdoğan (2012). Imputation with FHDI method was conducted with “FHDI” (Im et al.,2018) package.

## Results

Reference DIF results obtained from complete data set appear in Table 2. Those results were compared with DIF results of all combinations included in the study. We examined whether numbers, levels or directions of DIF items in complete data set have changed.

**Table 2**

*DIF Results for MH and MIMIC Methods in Complete Data set*

| | MH | | MIMIC |
|---|---|---|---|
| Item | $\Delta_{\mathrm{MH}}$ | Level | Beta |
| Item1 | 0.071 | - | 0.038 |
| Item2 | -1.422 | B (R) | -0.297* |
| Item3 | -0.226 | - | 0.027 |
| Item4 | 0.178 | - | 0.047 |
| Item5 | -0.815 | - | -0.267* |
| Item6 | -0.641 | - | -0.110 |
| Item7 | 0.491 | - | 0.089 |
| Item8 | 0.332 | - | 0.153* |
| Item9 | 0.326 | - | 0.097 |
| Item10 | -1.313 | B (R) | -0.218* |
| Item11 | -0.750 | A (R) | -0.178* |
| Item12 | 0.489 | - | 0.068 |
| Item13 | 0.221 | - | 0.126 |
| Item14 | 0.664 | - | 0.098 |
| Item15 | 0.732 | - | 0.113 |
| Item16 | 0.175 | - | 0.031 |
| Item17 | 1.367 | B (F) | 0.220* |

**Items showing DIF; R: Favors reference group F: Favors focal group*

As can be seen from the Table 2, two items displayed B level DIF and one item displayed A level DIF favoring reference group with MH method. One B level DIF item favoring focal group was also detected.

MIMIC method identified six DIF items. Four items (Item2, Item5, Item10 and Item11) favored reference group and two items (Item8 and Item17) favored focal group.

Table 3 illustrates DIF items with MH method in all combinations. DIF results were not reported for 15% missing condition with LD as it reduced sample size (70 students in total) dramatically. Sample size for 5% condition with LD was 457 students (232 students for the reference group and 225 students for the focal group).

**Table 3**

*DIF Results for MH Under MCAR Mechanism*

| Method | 5% | $\Delta_{\mathrm{MH}}$ | 15% | $\Delta_{\mathrm{MH}}$ |
|---|---|---|---|---|
| Zero Imputation | Item2**(R)<br>Item10*(R)<br>Item11*(R)<br>Item17**(F) | -1.271<br>-0.963<br>-0.774<br>1.054 | Item2*(R)<br>Item8*(F)<br>Item11*(R)<br>Item17*(F) | -0.698<br>0.712<br>-0.820<br>0.912 |
| Listwise Deletion | Item10***(R)<br>Item14**(F)<br>Item17***(F) | -1.863<br>1.298<br>2.155 | -------- | --------- |
| Fractional Hot-Deck | Item2***(R)<br>Item10***(R)<br>Item11*(R)<br>Item14**(F)<br>Item17***(F) | -1.658<br>-1.591<br>-0.933<br>1.017<br>1.500 | Item2**(R)<br>Item10*(R)<br>Item11**(R)<br>Item12*(F)<br>Item15**(F)<br>Item17**(F) | -1.344<br>-0.922<br>-1.158<br>0.623<br>1.210<br>1.191 |

**:Item showing A level DIF **:Item showing B level DIF ***:Item showing C level DIF Significance level:0.05*

As shown in Table 3, directions of DIF items in complete data set did not change in all conditions. However, there have been differences in number of DIF items and DIF magnitude. Three missing data methods produced following results for 5% condition. DIF items remained the same with ZI, but DIF magnitude of one item (item10) decreased. When LD was used, two DIF items (item10 and item 17) did not change except that they had higher DIF value than their actual value. Item14 displayed DIF with LD although it was not among DIF items in complete data set. Four DIF items were identified correctly with FHDI, yet three of them were overestimated. Item14 showed false DIF in favor of focal group.

When the missing data percentage was 15%, ZI and FHDI both obtained false DIF. ZI identified three of the four DIF items in complete data set while FHDI identified them all. DIF magnitude of two items (item2 and item17) were underestimated with ZI. FHDI produced overestimated DIF magnitude for item11 while it underestimated the DIF magnitude of item10. Table 4 presents DIF items with MIMIC method in all combinations.

**Table 4**

*DIF Results for MIMIC Under MCAR Mechanism*

| Method | 5% | Beta | 15% | Beta |
|---|---|---|---|---|
| Zero Imputation | Item2(R)<br>Item5(R)<br>Item10(R)<br>Item11(R)<br>Item17(F) | -0.278<br>-0.256<br>-0.178<br>-0.195<br>0.157 | Item8(F)<br>Item11(R)<br>Item17(F) | 0.145<br>-0.195<br>0.153 |

**Table 4**

*DIF Results for MIMIC Under MCAR Mechanism (Continued)*

| Method | 5% | Beta | 15% | Beta |
|---|---|---|---|---|
| Listwise Deletion | Item9(F) | 0.229 | | |
| | Item10(R) | -0.278 | --------- | -------- |
| | Item17(F) | 0.267 | | |
| | Item2(R) | -0.302 | Item2(R) | -0.196 |
| | Item5(R) | -0.293 | Item8(F) | 0.176 |
| | Item8(F) | 0.174 | Item10(R) | -0.139 |
| Fractional Hot-Deck | Item10(R) | -0.256 | Item11(R) | -0.281 |
| | Item11(R) | -0.185 | Item13(F) | 0.172 |
| | Item13(F) | 0.148 | Item15(F) | 0.200 |
| | Item14(F) | 0.197 | | |
| | Item17(F) | 0.228 | | |

**Items showing DIF; R: Favors reference group F: Favors focal group*

As in MH method directions of DIF items in complete data set did not change in all conditions for MIMIC method. Three missing data methods produced following results for 5% condition. ZI could not identify only one DIF item which had DIF in complete data set analysis. LD obtained false DIF for item9. Two out of six DIF items showing DIF in complete data set were determined as DIF items with LD.FHDI identified all DIF items accurately; however, it produced false DIF in favor of focal group in two items. In the case of 15% condition, both ZI and FHDI methods were unable to correctly identify all items indicating DIF in complete data set. Nevertheless, FHDI produced false DIF for this condition while ZI did not. Table 5 shows percentage of correctly identified DIF items and DIF free items by missing data handling methods with MH and MIMIC.

**Table 5**

*Percentage of Correctly Identified DIF Items and DIF Free Items by Missing Data Handling Methods*

| DIF Detection Method | Missing Data method | 5% | 15% |
|---|---|---|---|
| MH | | | |
| | ZI | 100% | 75% |
| *Percentage of correctly identified DIF items* | LD | 50% | ---- |
| | FHDI | 100% | 100% |
| *Percentage of correctly identified DIF free items* | ZI | 100% | 92% |
| | LD | 92% | ---- |
| | FHDI | 92% | 85% |
| MIMIC | | | |
| | ZI | 83% | 50% |
| *Percentage of correctly identified DIF items* | LD | 33% | ---- |
| | FHDI | 100% | 67% |
| *Percentage of correctly identified DIF free items* | ZI | 100% | 100% |
| | LD | 90% | ---- |
| | FHDI | 81% | 81% |

When examined in terms of the percentage of correctly identified DIF items and DIF free items in complete data set, it was found that for 5% condition with MH method, ZI and FHDI identified all DIF items in complete data set correctly. On the other hand, percentage of DIF items which were correctly identified by LD was 50%. DIF free items were the same with ZI. For this condition, 92% of DIF free items did not display DIF with LD and FHDI methods. FHDI determined all DIF items accurately for 15% missing case whereas percentage of DIF items obtained with ZI was 75%. Percentage of DIF free items which were correctly identified was 92% and 85% for ZI and FHDI methods respectively.

When MIMIC method was used, it was found that for 5% condition, FHDI identified all DIF items in complete data set accurately. Percentage of DIF items which were correctly identified were was 33% and 83% for LD and ZI respectively. Items that did not show DIF in complete data set were determined correctly with ZI. However, the percentage of DIF free items that were correctly identified by LD and FHDI were 90% and 81%.

FHDI was able to identify correctly 67% of DIF items for %15 missing case. The result was 50% for ZI in the same condition. ZI was better than FHDI in detecting DIF free items. ZI identified all DIF free items in complete data set correctly. On the other hand, the percentage of DIF free items correctly identified with FHDI was 81%. Table 6 provides correlations of MH and MIMIC DIF statistics in all conditions.

**Table 6**

*Correlations of MH and MIMIC DIF Statistics in All Conditions*

| DIF Method | Complete Data | ZI (5%) | LD (5%) | FHDI (5%) | ZI (15%) | FHDI (15%) |
|---|---|---|---|---|---|---|
| MH | | | | | | |
| *Complete data* | 1 | .975* | .826* | .985* | .825* | .913* |
| MIMIC | | | | | | |
| *Complete data* | 1 | .968* | .671* | .981* | .795* | .808* |

**Correlation is significant at the 0.01 level*

As Table 6 shows, all coefficients are positive and significant at p<.01. FHDI has the highest correlations for 5% and %15 missing case with both DIF methods. This result indicates that FHDI produces the closest DIF values to the values obtained from the complete data set. LD has the lowest correlation with both DIF methods. The correlations are slightly higher for MH method than MIMIC in all conditions.

## Discussion

This study was designed to examine the impact of missing data techniques (ZI, LD and FHDI) on performances of MH and MIMIC DIF detection methods under MCAR missing data mechanism. Missing data percentage was set at 5% and 15%. The current study found that the percentage of identifying DIF items with LD was quite low for both DIF detection methods. It also produced the lowest correlations with reference DIF values regardless of the DIF detection method used. When the missing data percentage increased, sample size was reduced considerably with LD which resulted in no clear DIF results and could not be reported. This limitation was also reported by Emenogu et al. (2010) who could not calculate all DIF statistics with LD in their research.

Another important finding was that for both DIF detection methods, FHDI was the best in identifying the percentage of DIF items in all conditions while ZI was more successful than the other two methods in finding DIF free items. In terms of the correlations between the DIF statistics obtained from complete data set and the other conditions, FHDI had the highest correlations meaning it had the closest DIF values to the nonresponse data. ZI produced slightly lower correlations than FHDI. As regards to DIF detection methods, the results of the study indicated that the correlations are slightly higher for MH method than MIMIC in all conditions which suggests MIMIC method was more adversely affected than MH by the presence of missing data.

In the present study, the percentage of correctly identified DIF items with ZI was lower for the cases with higher missing data percentage regardless of the DIF detection method. Finch (2011b) reported that power rates for ZI decreased as the percentage of missing data increased in the study investigating the impact of missing data on nonuniform DIF detection. The most obvious finding of the current study was that LD was the least optimal method for both identifying DIF items and DIF free items in complete data set. FHDI performed well in correctly identifying DIF items whereas ZI performed better than the other two methods in determining DIF free items in complete data set. In this case, the choice of missing data method should be based upon whether it is more essential to correctly identify items as DIF or falsely do so.

In this investigation, we aimed to study only with real data, which was a limitation of the research. There was only one sample size used in the study as we could not reach larger samples appropriate for our research. Relatively small sample size did not allow us to vary missing data rate; however, there might be missing data more than 15% in real life situations. Research is also needed to determine the performances of missing data handling methods (especially FHDI as it was the best of all) with larger samples and missing data rates.

As mentioned before there has been an increasing attention on DIF detection with MIMIC method. However, most studies in the literature have not dealt with DIF detection with CFA-based methods in detail when missing data is present. The aim of this study was to contribute to the literature on DIF detection with missing data by comparing two different methods based on CTT and CFA respectively. Since the study was limited to MH and MIMIC methods, it was not possible to see the performances of other methods based on CTT or IRT. Further work needs to be done to examine the performance of MIMIC method with missing data. Researchers might explore the effect of sample size, DIF magnitude and other missing data treatment methods on DIF detection with MIMIC and compare those results with DIF detection methods other than MH.

## Declarations

**Author Contribution:** Rabia Akcan-Conceptualization, investigation, methodology, analysis, writing & editing. Kübra Atalay Kabasakal- Conceptualization, investigation, methodology, analysis, writing & editing, supervision.

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

**Ethical Approval:** Secondary data was used in this study.Therefore, ethical approval is not required.

## References

Banks, K. (2015). An introduction to missing data in the context of differential item functioning. *Practical Assessment, Research & Evaluation*, 20(12), 1-10. https://eric.ed.gov/?id=EJ1059748

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

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. London Sage.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement Issues and Practice*, 17(1), 31-44. https://eric.ed.gov/?id=EJ564712

Doğanay Erdoğan, B. (2012). Çoklu atama yöntemlerinin Rasch modelleri için performansının benzetim çalışması ile incelenmesi [Assessing the performance of multiple imputation techniques for Rasch models with a simulation study] (Publication No. 314412) [Doctoral dissertation, Ankara University]. Council of Higher Education Thesis Center.

Dong, Y., & Peng, C. Y. (2013). Principled missing data methods for researchers. *Springer Plus*, 2(1), 222. https://doi.org/10.1186/2193-1801-2-222

Emenogu, B. C., Falenchuk, 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

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(4), 281-301. https://doi.org/10.1080/08957347.2011.607054

Finch, H. (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

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement*, 29(4), 278-295. https://doi.org/10.1177/0146621605275728

Finch, H. W., & French, B. F. (2007). Detection of crossing differential item functioning a comparison of four methods. *Educational and Psychological Measurement*, 67(4), 565-582. https://doi.org/10.1177/0013164406296975

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education* McGraw-hill.

Garrett, P. L. (2009). *A monte carlo study investigating missing data, differential item functioning, and effect size* (Publication No. 3401601) [Doctoral dissertation, Georgia State University]. ProQuest Dissertations Publishing.

Hallquist, M., & Wiley, J. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(4), 621-638. https://doi.org/10.1080/10705511.2017.1402334

Harrington, D. (2009). *Confirmatory factor analysis*. Oxford University Press.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer, & H. I. Braun, *Test Validity* (pp. 129-145). Lawrence Erlbaum.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Lawrence Erlbaum.

Im, J., Cho, I. H., & Kim, J. K. (2018). *FHDI: Fractional hot deck and fully efficient fractional imputation*. https://CRAN.R-project.org/package=FHDI

Im, J., Kim, J. K., & Fuller, W. A. (2015). Two-phase sampling approach to fractional hot deck imputation. *In Proceedings of the Survey Research Methods Section*, pages 1030-1043. http://www.asasrms.org/Proceedings/y2015/files/233957.pdf

Jin, KY., & Chen, HF. (2020). MIMIC approach to assessing differential item functioning with control of extreme response style. *Behavior Research Methods*, 52, 23-35. https://doi.org/10.3758/s13428-019-01198-1

Kalton, G., & Kish, L. (1984). Some efficient random imputation methods. *Communications in Statistics-Theory and Methods* ,13(16), 1919-1939. https://doi.org/10.1080/03610928408828805

Kim, J. K., & Fuller, W. (2004). Fractional hot deck imputation. *Biometrika*, 91(3), 559-578. https://doi.org/10.1093/biomet/91.3.559

Magis, D., Beland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, 42(3), 847-862. https://doi.org/10.3758/BRM.42.3.847

Montoya, A. K., & Jeon, M. (2020). MIMIC models for uniform and nonuniform DIF as moderated mediation models. *Applied Psychological Measurement*, 44(2), 118-136. https://doi.org/10.1177/0146621619835496

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research:A review of reporting practices and suggestions for improvement. *Review of Educational Research*, 74(4), 525-556. https://journals.sagepub.com/doi/pdf/10.3102/00346543074004525

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

Rousseau, M., Bertrand, R., & Boiteau, N. (2004, April). *Impact of missing data on robustness of DIF IRT-based and non IRT-based methods*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581-592. https://doi.org/10.1093/biomet/63.3.581

Scheuneman, J. (1979). A method of assessing bias in test items. *Journal of Educational Measurement*, 16, 143–152. https://www.jstor.org/stable/1433816

Shih, C. L., & Wang, W. C. (2009). Differential item functioning detection using multiple indicators, multiple causes method with a pure short anchor. *Applied Psychological Measurement*, 33(3), 184-199. https://doi.org/10.1177/0146621608321758

Tamcı, P. (2018). Kayıp veriyle baş etme yöntemlerinin değişen madde fonksiyonu üzerindeki etkisinin incelenmesi [Investigation of the impact of techniques of handling missing data on differential item functioning] (Publication No. 517260) [Master's dissertation, Hacettepe University]. Council of Higher Education Thesis Center.

Uğurlu, S., & Atar, B. (2020). Performances of MIMIC and logistic regression procedures in detecting DIF. *Journal of Measurement and Evaluation in Education and Psychology*, 11(1), 1-12. https://doi.org/10.21031/epod.531509

Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research*, 44(1), 1-27. https://doi.org/10.1080/00273170802620121

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland, & H. Wainer, *Differential Item Functioning* (pp. 337-347). Lawrence Erlbaum.

Zumbo, B. D. (2007). Three generation of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*(2), 223–233. https://doi.org/10.1080/15434300701375832