ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(4); 408-420

The Effect of Missing Data Handling Methods on Differential Item Functioning with Testlet Data*

Rabia AKCAN ** Kübra ATALAY KABASAKAL ***

Abstract

This study examined the effect of three missing data handling methods (listwise deletion, zero imputation and fractional hot-deck imputation) on differential item functioning (DIF) with testlet data with a variety of sample size and missing data percentage under missing completely at random, missing at random, and missing not at random missing mechanisms. The study was conducted on two different datasets consisting of six testlets which contain 20 reading comprehension items of a foreign language test. Data with left-skewed distribution was referred to as data1 and data with right-skewed distribution was referred to as data2. In current study, false DIF was identified in data1 with all missing data methods under the missing at random mechanism with a 5% missing data rate in small sample size. Similarly, in analyses performed under the missing at random mechanism for data2, the proportion of items classified as false DIF was notably higher in the small sample size. Results also indicated that in all conditions, list wise deletion had the lowest correlations with DIF values obtained from the original datasets, datasets containing no missing data and serve as a reference for comparative analyses with datasets where missing data were artificially introduced. The zero imputation and fractional hot-deck imputation methods produced similar correlations when the missing data percentage was set at 5%. However, in the case of 15% missing data, zero imputation exhibited higher correlation values. Besides, in all conditions correlation values decreased with the increase of missing data percentage regardless of the missing data handling method.

Keywords:testlet, missing data, differential item functioning

Introduction

To date, there has been an ongoing investigation on defining and achieving validity. Validity can be defined as the degree to which evidence and theory support the test scores' interpretations for intended uses of tests. It is the most essential consideration for the development and evaluation of tests (AERA et al., 2014). Therefore, accumulating evidence for the validity of test scores is crucial for effective test development and evaluation.

Item bias is one of the key issues in test validity. It becomes evident when examinees of one group have a lower probability of success on the item than examinees of another group at the same ability level due to some characteristic of the test item or testing situation that is irrelevant to the test purpose (Zumbo, 1999). If any test item provides advantage to one of the groups, it negatively affects validity. Therefore, bias studies can play an important role in addressing the issue of validity.

The first step in identifying item bias is to detect items containing differential item functioning (DIF). DIF occurs when examinees from different groups, who have been matched on the ability of interest, have differing probabilities or likelihoods of succeeding on an item (Clauser & Mazor, 1998). DIF is

*This study was produced from the doctoral dissertation conducted by the first author under the supervision of the second author.

**Teacher., Republic of Turkey Ministry of National Education, Afyonkarahisar-Turkey, eltrabia42@hotmail.com, ORCID ID: 0000-0003-3025-774X
***Assoc. Prof., Hacettepe University, Faculty of Education, Ankara-Turkey, katalay@hacettepe.edu.tr, ORCID ID: 0000-0002-3580-5568

To cite this article:
Akcan, R., Atalay Kabasakal, K., (2024). The effect of missing data handling methods on differential item functioning with testlet data. *Journal of Measurement and Evaluation in Education and Psychology*, *15*(4), 408-420. https://doi.org/10.21031/epod.1539940

Received: 28.08.2024
Accepted: 24.11.2024

required, but not sufficient condition for item bias. If DIF is present, follow-up item bias analyses (e.g., content analysis) would have to be done to determine whether item bias is evident or not (Zumbo, 1999). In other words, DIF is a statistical technique which helps identifying potentially biased items.

DIF analyses play a vital role in test development and validation to assure that scores obtained from educational tests and psychological measures are not biased and they measure the same construct for all examinees (Walker, 2011). Although standalone item DIF analysis has received considerable attention, small bundles of items are the fundamental building blocks of many exams. A bundle is defined as any group of items chosen in accordance with some organizing principle. These items do not have to be adjacent or it is not necessary for them to refer to a common passage or a text (Douglas et al., 1996). For example, three independent math items based on analytical reasoning can form an item bundle in a test.

DIF analyses can be carried out both at the item and bundle level. Item bundle DIF, which is called as differential bundle functioning (DBF), is an extended form of item DIF (Douglas et al., 1996). As previously stated, items in a bundle are not necessarily close to each other or they do not have to share a common passage. However, Beretvas and Walker (2012) point out that there are various reasons to put the items together in a bundle. One of them is the testlets in which items might be bundled together. A testlet is a set of items which are based on a common stimulus (Wainer & Kiely, 1987). For instance, items within a testlet may focus on a laboratory scenario, a graphic, a reading passage or complex problem (DeMars, 2006). Testlets save testing time since examinees focus on the scenario once and they can utilize the information for other items. Besides, authenticity of the task may increase as more context is added (DeMars, 2012). A well-known example of testlets is reading comprehension items which are based on a paragraph in language tests. The difference between a testlet and an item bundle is that items in a testlet share a common input whereas this is not the case for an item bundle. Therefore, examining item bias in testlets with a different method provides more valid and reliable results.

SIBTEST (Shealy & Stout, 1993) and Poly-SIBTEST (Chang et al., 1996), which is an extended form of SIBTEST for polytomous items, have been commonly used in the detection of DIF at the item level and DBF at the testlet level (Beretvas & Walker, 2012; Lee, Cohen & Toro, 2009; Min & He, 2020). DIF can only be identified at the testlet level when DBF analysis is carried out within the framework of SIBTEST method, proposed by Douglas et al. (1996), and thus may be referred as differential testlet functioning. In this case, it is not possible to determine which items are causing differential testlet functioning. Moreover, creating a testlet is expensive and time-consuming. It's better to handle problems at the item level by determining problematic items instead of discarding the whole testlet from the item bank due to differential testlet functioning. Accordingly, it would be more practical and useful to apply a method which examines DIF at the item level rather than a method examining differential testlet functioning (Fukuhara & Kamata, 2011).

A Bifactor MIRT Model ForTestletsWith Covariates

Fukuhara and Kamata (2011) proposed a DIF detection model which is an extension of a bifactor multidimensional item response theory (MIRT) model for testlets. Unlike conventional item response theory (IRT) DIF models, this proposed model takes testlet effects into consideration. Consequently, it estimates DIF magnitude appropriately if a test consists of testlets. Moreover, DIF can be identified for all items simultaneously with the proposed DIF model. It also estimates DIF magnitudes assuming that the average DIF magnitude is zero. Besides, there is a parameter to capture the mean ability difference between the focal and reference groups to distinguish DIF and impact. If this parameter is not included, it is assumed that no ability difference between the focal and reference groups exists. The bifactor MIRT model for testlets with covariates proposed by Fukuhara and Kamata (2011) is reduced to a traditional IRT model if there is no testlet effect, which ensures that there will be no adverse impact when the testlet effect is not present. The authors set the absolute value of 0.426 in the logit scale as the threshold for a meaningful DIF magnitude in their study. The current study utilized

the proposed model by Fukuhara and Kamata (2011) for the DIF detection process and the absolute value of 0.426 was adopted to identify meaningfully large DIF items as the researchers did in their study.

Missing Data

Another significant aspect of validity is missing data which may cause incorrect trait inferences, thereby reducing validity (Garrett, 2009). Any blank responses to the entire set of items that a test taker has access to are referred to as missing data (Ludlow & O'leary, 1999). In real life assessment scenarios, missing data are widely seen. There are various methods to handle missing data. Type of missing data and the method chosen to handle missing data can have a unique impact on the statistical results (Garrett, 2009).

In order to decide on the appropriate method to handle missing data, one must address missing data type. Rubin (1976) classified missing data mechanisms into three types: Missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Data are MCAR when there is no relation between the probability of missing data on a variable Y and the value of Y itself or the values of any other variables in the dataset (Allison, 2002). MAR data is present if the probability of missing data on the variable Y is related to any other variables in the model but it is not related to the values of Y itself (Enders, 2010). In other words, data are MAR when the probability of missing data is just influenced by the values of other observed variables (Robitzsch & Rupp, 2009). Data are MNAR if the probability of missingness on the variable Y is related to the values of Y itself, even after controlling for other observed variables (Enders, 2010). Missingness cannot be explained with observed variables for MNAR data, as it depends on the unobserved values (Robitzsch & Rupp, 2009).

In real life situations DIF and missing data might occur at the same time. It is essential to investigate DIF in the presence of missing data. Nevertheless, commonly used DIF detection methods such as Mantel Haenszel (MH), SIBTEST and Logistic Regression (LR) cannot handle missing data (Banks, 2015). To be more specific, In MH DIF analysis for example, students' abilities are typically matched based on the number of items they answer correctly. Since this matching is based on the number correct answers, missing items are generally considered as either not administered or incorrectly answered. Thus, it is crucial to investigate how these methods and other missing data handling methods affect DIF analysis (Emenogu et al., 2010).

Missing data handling methods applied for the analysis may also lead to 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). There have been research on DIF detection in the presence of missing data with simulated data (Finch, 2011a; Finch, 2011b; Garrett, 2009; Robitzsch & Rupp, 2009) or real data (Rousseau et al., 2004; Tamcı, 2018). Most of these studies have focused on DIF and missing data in different aspects. However, very little attention has been paid to the role of missing data on DIF detection with testlet data.

Since items in a testlet are dependent, when an examinee leaves an item blank in the testlet, responses to other items will probably be affected, which creates MNAR data. If the examinees from one group leave the items blank at a larger rate than those from the other group, this produces unbalanced data (Sedivy, 2009). As a result, it is important to investigate the impact of missing data and missing data handling methods on DIF detection with testlet data.

A number of techniques have been developed to handle missing data problem. The present study used listwise deletion (LD), zero imputation (ZI) and fractional hot-deck imputation (FHDI) to deal with missing data. In LD, any observations with missing data on the variables are deleted from the sample. One advantage of LD is that it can be applied for any type of statistical analysis. Another advantage of it is that it does not require any special computational technique (Allison, 2002). However, there are certain drawbacks associated with the use of LD. Primarily, it requires MCAR data and may create inaccurate parameter estimates when this assumption is ignored. It can also produce biased estimates when the data are only MAR, but not MCAR. Aside from bias, if discarded observations have data on

many variables, discarding observations with missing data reduces total sample size drastically and thus causes decrease in statistical power (Allison, 2002; Enders, 2010). In literature, however, LD has been widely used in research investigating missing data and DIF together (Banks & Walker, 2006; Emenogu et al., 2010; Finch, 2011a; Finch, 2011b; Robitzsch & Rupp, 2009; Sedivy et al., 2006). As already stated, in real life situations dependency of the items in a testlet is likely to cause MNAR data if an examinee leaves an item blank, so it would be interesting to see how LD method works under three different missing data mechanisms on DIF detection with testlet data.

ZI is one of the most basic techniques for imputing item response data among the techniques that employ a single imputation step. In this method, all missing values are replaced with a score of zero. Nonetheless, some researchers do not consider this method as a true imputation method because it lacks a statistical model. However, it is frequently used in the context of achievement tests because it is easy to do and can reasonably suggest that a lack of response shows lack of proficiency (Robitzsch & Rupp, 2009).

FHDI, proposed by Kalton and Kish (1984) and investigated by Kim and Fuller (2004), is a way of performing hot deck imputation efficiently. In this method, M imputed values are produced for each missing value as in multiple imputation (MI); nevertheless, a single dataset is obtained as the output after fractional imputation (Im et al., 2015). The imputed values are randomly selected from the donors' data within the same imputation cell. These cells are particularly created to ensure data homogeneity within each cell. Fractional weights are assigned to each imputed value to maintain the original data structure and for variance estimation replication methods are adopted (Im et al., 2015; lm et al., 2018). FHDI was extended by Im et al. (2015) in two ways. First, in this new version of FHDI, a nonparametric imputation approach, imputation cells are not required to be made in advance. Instead, multiple cells are allowed for each missing item. Second, the proposed FHDI method is applied to multivariate missing data with arbitrary missing patterns. In the current study, we utilized extension of FHDI proposed by Im et al. (2015) as the imputation method. Fractional imputation has not yet seen widespread adoption outside of survey sampling, probably because it is a relatively new method and involves more complex variance estimation procedures compared to MI (lm et al., 2018). However, fractional imputation provides consistent variance estimates, especially when using a method-of-moment estimator. In contrast, MI may sometimes yield inconsistent variance estimates (Yang and Kim, 2016). FHDI approach used in this study utilizes observed values as imputed values, and thus can better preserve the structure of the data. Despite the widespread use of MI, only one of the two methods was employed in this study due to practical constraints-specifically the extended analysis time required for DIF analyses. FHDI was selected for this study due to its noted advantages and its potential as a promising method.

Research on DIF in the presence of missing data has produced various results: In a simulation study Finch (2011a) found that compared with LD and MI, ZI had highly inflated type I error rates under MAR mechanism and it was found to be the least applicable method under this condition. Results also indicated that LD and MI performed similarly. Another simulation study by Finch (2011b) also demonstrated that LD was superior to ZI under various conditions. In their simulation study, Robitzsch and Rupp (2009) concluded that incorrect choice of missing data method led to false DIF. In addition, missing data handling methods had less problematic results under MCAR mechanism. Several studies used real data to investigate the impact of missing data handling methods on DIF detection. Akcan and Atalay Kabasakal (2023) focused on the impact of missing data on DIF detection using LD, ZI and FHDI methods under MCAR mechanism. They reported that FHDI yielded the best results in detecting DIF items in all conditions and DIF values obtained with FHDI were the closest DIF values to those obtained from the original dataset. Tamcı's (2018) study on DIF detection with a variety of DIF magnitude, sample size and focal/reference group rate in case of MCAR data showed that MI had lower error rates than ZI and expectation-maximization. Power rates for these methods were mostly below the acceptable level with the exception of ZI for 10% missing data percentage. Emenogu et al. (2010) investigated the impact of ZI, LD and analysis wise deletion on MH method with both real and simulated data. 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.

Nichols et al. (2022) assessed DIF on a real dataset by using single hot-deck imputation and Multiple Imputation by Chained Equations (MICE) as a multiple imputation technique with a variety of missing rate and DIF scenarios. They reported that MICE achieved slightly better results than hot deck single imputation in reducing observed DIF estimation errors, although both methods were effective in decreasing observed errors compared to scenarios without any imputation. They suggested using MICE in testing DIF to reduce the bias caused by missing data when the missing data rate exceeds the 10% threshold. They also stated that MICE could not remove the observed error due to missing data in their study. As a result, they advised investigators to interpret results with caution when they employ MICE to handle missing cognitive data.

Purpose of the Study

Testlets are widely utilized in many high-stakes testing situations (e.g. American College Testing, Graduate Record Examination, Test of English as a Foreign Language and International English Language Testing System). Various studies have investigated DIF in testlet-based items to determine the effect of testlets on DIF detection (Fukuhara & Kamata, 2011; Min & He, 2020; Ravand, 2015; Sedivy, 2009; Taşdelen Teker, 2014; Wang & Wilson, 2005). It is inevitable that there will be missing data in real life situations. Although testlets are widely used in large scale examinations, there have been no attempts to investigate the effect of missing data handling methods on DIF with testlet data. Determining the conditions that missing data handling methods work best will contribute to the accuracy of DIF detection results. Therefore, this study provides new insights into DIF detection with testlets in the presence of missing data. The leading research question in this investigation is as follows: Do missing data handling methods have an impact on DIF detection with testlet data with a variety of sample size and missing data percentage? To achieve the goal of this research, following sub-problems are addressed:

1) How do DIF results change across sample size (1,000 and 2,000) and missing data percentage (5% and 15%) under MCAR, MAR and MNAR mechanisms by using LD, ZI and FHDI missing data handling methods?

2) How do the correlations between DIF magnitudes obtained from the original datasets and new datasets change across sample size (1,000 and 2,000) and missing data percentage (5% and 15%) under MCAR, MAR and MNAR mechanisms by using LD, ZI and FHDI missing data handling methods?

Method

Dataset

The dataset used in this research was obtained from students' responds to six testlets composed of 20 items in English test of Undergraduate Placement Exam (UPE) conducted in Turkey in 2016. Items that formed testlets were cloze test (Items 1-5 where the participants have to fill in the blanks in one reading passage), reading-1 (Items 6-8), reading-2 (Items 9-11), reading-3 (Items 12-14), reading-4 (Items 15-17) and reading-5 (Items 18-20). To get a complete dataset composed of testlets, all examinees having missing values were deleted from testlet data and a dataset consisted of 33570 examinees was created. Four schools out of 87 were chosen as the sample because they had sufficient sample size for the purpose of this study. These schools were private high schools teaching in foreign language (4275 students), science high schools (1183 students), religious high schools (2333 students) and formal high schools (4891).Two different datasets were created from these schools according to their distributions. Data distribution was regarded as another condition. Data1 consisted of private high schools teaching in foreign language and science high schools which had left-skewed distributions. Data2, on the other hand, consisted of religious high schools and formal high schools which had right-skewed distributions. Two sample size conditions (1,000 and 2,000) were included in

the study which accounted for four samples in total. These four samples were referred as original datasets.

Data Analysis

Datasets used in the study were data1 and data2, which had left-skewed and right-skewed distributions, respectively. Study was conducted on four samples (1,000 and 2,000 sample size for each) which were drawn from these datasets. To begin the process, DIF analyses were performed on four samples by using the bifactor MIRT model for testlets with covariates and results were used as reference. Missing data were generated under MCAR, MAR and MNAR mechanisms and percentage of missing data was set at 5% and 15%, and thus 24 datasets including missing data were created from the four original samples. Following this missing data generation process, LD, ZI and FHDI were adopted to handle missing data problem of 24 datasets and 72 datasets without missing data problem were obtained. Finally, DIF analyses were conducted on 72 datasets by using the bifactor MIRT model for testlets with covariates to compare the results with those obtained from the original datasets in the first stage of the process. LD and ZI were performed by writing codes on base R and "FHDI" (lm et al., 2018) package was used for imputation with FHDI. Testlet DIF analyses for all datasets were carried out using WinBUGS 1.4.3 (Spiegelhalter et al., 2003).

Missing Data Generation Process

Missing data were generated on the items under MCAR, MAR and MNAR mechanisms. Percentage of missing data was set at 5% and 15% for the entire dataset. In case of MCAR data, appropriate proportion of responses on all items from both reference and focal groups were randomly selected. For MAR data, responses on all items were randomly selected only from the focal group. Percentage of missing data deleted from the focal group in MAR case was set at 5% and 15% of the entire dataset. Under MNAR mechanism, missing data were generated in both groups and it depended on the item difficulties and total scores. Yet, percentage of missing data was set at the same proportion as MCAR and MAR data for the entire dataset. In MNAR missing data generation process, total scores were divided into three levels from lowest to the highest whereas item difficulties were divided into three levels from easiest items to the most difficult ones. Certain percentages of data were deleted in each level, total amount of which was equal to 5% or 15% of the entire dataset. To clarify, the amount of missing data was greater for the examinees with the lowest ability levels on the most difficult items compared to the examinees with the highest ability levels and so on. Missing data were created in R software by adapting the codes written by Doğanay Erdoğan (2012) to this study.

DIF Analyses

DIF analyses for all datasets were carried out using the bifactor MIRT model for testlets with covariates. Parameters were estimated using WinBUGS 1.4 program with a Markov chain Monte Carlo (MCMC) method. When the MCMC method is used to estimate parameters, it should be checked whether the parameter estimates converge. If convergence of the parameters is not achieved, incorrect inferences regarding the parameter of interest will be drawn. Therefore, it is necessary to decide the number of iterations to remove (i.e., burn-in iterations) when a parameter estimation becomes stable. Besides, the numbers of iterations after a burn-in period needs to be decided to get good samples of each parameter that represent the parameter's posterior distribution (Fukuhara & Kamata, 2011). In this study, preliminary analyses were conducted on the original datasets, and the burn-in iterations and total number of iterations were determined to achieve convergence. Based on the preliminary analysis, 9,800 samples were drawn from each posterior distribution after discarding 200 samples as burn-in period. Fukuhara and Kamata (2011) assessed convergence using history plots, density plots and auto-correlation plots in their study. Another way of assessing convergence is to check MC errors. Small values of MC errors show that parameter of interest is estimated accurately

(Ntzoufras, 2009). To assess convergence the present study used graphical methods (history, density and autocorrelation plots) and also checked MC errors. Results indicated that the parameter estimates of DIF converged.

Results

The first set of analyses determined the DIF items in original datasets. The bifactor MIRT model for testlets with covariates proposed by Fukuhara and Kamata (2011) was used to identify DIF and the value of 0.426 was adopted to identify meaningfully large DIF items as the researchers did in their study. Table 1 presents DIF items in each original sample.

Table 1
DIF results in original datasets.

	Data1		Data2	
Item No	**1000**	**2000**	**1000**	**2000**
1	0.080	0.016	0.393	**0.455***
2	0.044	-0.115	-0.178	-0.084
3	**-0.474***	**-0.484***	-0.326	-0.260
4	0.164	0.006	**0.439***	**0.480***
5	0.071	0.025	-0.051	-0.079
6	-0.028	-0.052	-0.336	-0.105
7	-0.149	-0.105	-0.379	-0.175
8	-0.032	-0.007	-0.246	-0.184
9	-0.080	0.058	0.270	0.048
10	-0.129	-0.113	0.051	0.093
11	-0.184	-0.081	-0.161	-0.106
12	-0.024	-0.076	**-0.526***	-0.217
13	0.174	0.120	0.290	0.198
14	0.067	0.125	-0.059	-0.113
15	0.049	0.052	0.148	0.073
16	0.382	0.235	0.424	**0.435***
17	-0.056	0.047	-0.078	-0.172
18	0.194	0.173	0.209	-0.201
19	-0.153	0.001	-0.099	-0.171
20	0.084	0.177	0.214	0.085

**DIF magnitude>0.426*

It is apparent from Table 1 that only item 3 showed DIF in both sample sizes in data1. The other four items (items 1, 2, 4 and 5) in the same testlet with item 3 did not display significant DIF and their magnitudes were quite low. Results obtained from data2 indicated that only one DIF item (item 4) was common in the two sample sizes. Two items (items 4 and 12) were flagged DIF in small sample whereas three items (items 1, 4 and 16) showed DIF in the larger sample. In addition, some non-DIF items (items 1, 7 and 16) in small sample size had DIF magnitude relatively close to the value of 0.426.

Table 2 provides DIF items in all conditions for data1 after the treatment with ZI, LD and FHDI methods. At the beginning there were 36 conditions in total, however, eight of them could not be carried out because of the reduced sample size with LD method and presence of missing data in all cases for the focal group with FHDI method.

Table 2
DIF items in all conditions for data1 with ZI, LD and FHDI methods.

Method	Item No.					
	5%			15%		
	MCAR	**MAR**	**MNAR**	**MCAR**	**MAR**	**MNAR**
1000						
ZI	3	3,16	3	3	1, 3, 4,16	3
LD	*	16	3	-	-	-
FHDI	3	3,16	3	3,16	-	3, 5,18
2000						
ZI	3	3	3	3	3, 4,16	3
LD	3	3,20	*	-	-	-
FHDI	3	3	3	3,18	-	3, 10,16

**None of the items displayed DIF.*

What stands out in Table 2 is ZI and FHDI methods were superior to LD in detecting DIF item (item3) when the percent of missing data was set 5%. It was also found that the effects of three missing data handling methods on the performance of identifying DIF free items were similar. Another important finding was that under MAR mechanism in small sample size case, item16 showed false DIF with the three missing data handling methods.

For 15% missing case, only ZI and FHDI results were compared. As can be seen from the table, item3 was correctly identified as DIF item in all conditions. However, under MCAR and MNAR mechanisms, ZI performed better than FHDI method in terms of identifying DIF free items in the original datasets. On the other hand, ZI produced false DIF under MAR mechanism and the percentage of items that showed false DIF with ZI was found to be higher in smaller sample size. Closer inspection of the table shows that two items (item1 and item3) displaying false DIF in small sample under MAR mechanism with ZI in 15% missing case were in the same testlet. Likewise, two items (item3 and item5) displaying false DIF in small sample under MNAR mechanism with FHDI in 15% missing case were in the same testlet.

DIF items in all conditions for data2 after the treatment with ZI, LD and FHDI methods are presented in Table 3. At the beginning there were 36 conditions in total, however, six of them could not be carried out because of the reduced sample size with LD method and presence of missing data in all cases for the focal group with FHDI method.

Table 3
DIF items in all conditions for data2 with ZI, LD and FHDI methods.

Method	Item No.					
	5%			15%		
	MCAR	**MAR**	**MNAR**	**MCAR**	**MAR**	**MNAR**
1000						
ZI	4,12	7, 12,13	4,12	1, 4, 7, 12,16	6, 7, 10, 12, 13,16	*
LD	1, 2, 7, 12, 13,18	8,12	*	-	-	8, 15, 20
FHDI	4,12	1,12	12	1, 4, 6, 8,12, 13, 16, 18, 19,20	-	*
2000						
ZI	1,4	*	1,4	1, 4,16	4, 10, 13,16	1, 4
LD	4,20	*	*	-	-	4, 10
FHDI	1,4	4	4,16	1, 2, 4,16	-	4

**None of the items displayed DIF.*

From this data presented in Table 3, we can see that the percentage of items which had false DIF in small sample size was greater than or equal to those from the large sample size regardless of the missing data rate. Particularly, in analyses performed under the MAR mechanism, the proportion of items classified as false DIF is notably higher in the small sample size. As regards to the detection of DIF items in the original datasets, the results demonstrated that performances of ZI and FHDI were similar. Yet, these two methods led to false DIF in some conditions for both missing data percentage. Of interest here is the increase in error rate of classifying DIF and non-DIF items when the percentage of missing data was set 15%. Results also indicated that when the percentage of missing data was set 5% in case of MCAR and MNAR data, LD had the lowest performance on detecting DIF items regardless of the sample size. On the other hand, performance of identifying DIF free items in eight conditions improved with the increase in sample size.

It is apparent from Table 3 that some of the items that displayed false DIF (e.g. 18-20) were within the same testlets. It was also found that with ZI method, items that are in the same testlet and displayed false DIF together were mostly observed in case of MAR data. Table 4 shows the Pearson correlations between the DIF magnitudes obtained from all conditions and the ones obtained from the original datasets.

Table 4

Correlations between the DIF magnitudes obtained from all conditions and the original datasets.

Method		**5%**			**15%**	
	MCAR	**MAR**	**MNAR**	**MCAR**	**MAR**	**MNAR**
Data1_1000						
ZI	0.92**	0.91**	0.97**	0.89**	0.80**	0.85**
LD	0.83**	0.74**	0.60**	-	-	-
FHDI	0.98**	0.99**	0.97**	0.82**	-	0.72**
Data1_2000						
ZI	0.98**	0.90**	0.98**	0.95**	0.80**	0.94**
LD	0.88**	0.83**	0.73**	-	-	-
FHDI	0.97**	0.98**	0.89**	0.74**	-	0.77**
Data2_1000						
ZI	0.99**	0.97**	0.99**	0.97**	0.84**	0.93**
LD	0.88**	0.52*	0.90**	-	-	0.20
FHDI	0.98**	0.97**	0.99**	0.73**	-	0.92**
Data2_2000						
ZI	0.99**	0.95**	0.99**	0.97**	0.76**	0.96**
LD	0.71**	0.83**	0.64**	-	-	0.58**
FHDI	0.98**	0.98**	0.99**	0.68**	-	0.96**

**p<.05; **p<.01.*

From this data, we can see that LD method resulted in the lowest correlations in all conditions for both data1 and data2. Furthermore, in all conditions for both datasets there was a decrease in correlation values as the percentage of missing data increased. ZI and FHDI had similar results when the missing data percentage was set 5%. However, for 15% missing case ZI method had higher correlation values. Lowest correlation values with ZI were obtained under MAR mechanism.

The correlation results for data1 showed that when the sample size increased, a higher correlation was obtained with the LD method. Yet, a similar pattern was not obtained from the results of data2. It was determined that except for one condition, the correlation values obtained from data2 with the ZI method were higher than the values obtained from data1, which was not the case for LD and FHDI methods. The reason for this might be that data2 was skewed to the right, and thus DIF magnitudes were less affected by the imputation with ZI method.

Discussion and Conclusion

This study set out to investigate the impact of missing data handling methods on DIF detection with testlet data with a variety of sample size and missing data percentage. Study was conducted on four samples (1,000 and 2,000 sample size for each) which were drawn from two different datasets. For LD method, discarding observations with 15% missing data led to a significantly reduced sample size in most of the conditions, and thus DIF parameters could not be estimated with LD method under this condition. The study was limited to the results obtained with lower percentage of missing data with LD method. This finding is consistent with that of Emenogu et al. (2010).

Results obtained from both datasets showed that LD method produced the lowest correlations with reference DIF values in all conditions. While LD method performed as efficiently as or slightly lower than ZI and FHDI in detecting DIF free items, ZI and FHDI were superior to LD in detecting DIF items in many conditions for both datasets. This finding seems to be consistent with the research by Akcan and Atalay Kabasakal (2023).

Results from data1 indicated that in all conditions, there was an increase in error in detecting DIF free items with FHDI method as the percentage of missing data increased. It was also revealed that performance of ZI method in detecting DIF free items under MAR mechanism was adversely affected by the increase in missing data percentage. Likewise, there were conditions in data2, in which performance of detecting DIF free items decreased with the larger missing data percentage in three missing data mechanisms. In addition, correlation values obtained from the three missing data handling methods in all conditions decreased, as the percentage of missing data increased. These results are in agreement with those obtained by Emenogu et al. (2010). They stated in their research that impact of missing data handling method was insignificant when the missing data percentage was low. However, percentage of missing data in focal or reference group might be a source of DIF when the percentage of missing data was large.

Finch (2011a) reported that type I error rate of ZI method was greatly inflated in all conditions under MAR mechanism, which was not the case for LD and MI methods. The present study found that performance of identifying DIF free items under MAR mechanism was lower than MCAR and MNAR mechanisms in all conditions except for the results obtained from 2,000 sample size from data1 with ZI and FHDI methods. The two researches had similarities in this respect.

According to the correlations with reference DIF values, ZI and FHDI produced similar results in both datasets when the missing data percentage was set at 5%. On the other hand, correlations with ZI were higher than FHDI in case of 15% missing data. In their research, Akcan and Atalay Kabasakal (2023) reported that FHDI had the highest correlation values in all conditions and ZI had slightly lower correlation values than FHDI. This differs from the results presented here. A possible explanation for this might be the different distributions of the data used in these studies. Especially right skewed distributed datasets are likely to be less affected by imputation with ZI and produce closer DIF values to the values obtained from the original datasets.

Another finding was that performance of identifying DIF free items in eight conditions in data2 improved with the increase in sample size. This study also found that under MAR mechanism, there was a decrease in error in detecting DIF free items with ZI and FHDI methods when the sample size increased. These results are in agreement with Tamcı (2018) who showed that ZI yielded unacceptable type I error rates when the sample size decreased.

This current study was limited by two sample size as datasets used here were drawn from a real dataset. Missing data percentage was set at 5% and 15% due to the relatively small sample sizes. Nevertheless, it was not possible to estimate DIF parameters in some conditions with 15% missing data case because of the reduced sample size with LD method and presence of missing data in all cases for the focal group with FHDI method. The study did not include any other DIF detection methods for testlets. In future research, it would be beneficial to employ multiple DIF detection methods so that researchers can gain insights into how various DIF detection methods work in presence of missing data. Further research might also explore the impact of different missing data handling methods on

testlet DIF with a larger sample size and missing data percentage. The study can be repeated using both real data and simulated data using different DIF detection methods and missing data handling methods. A further study might also investigate DIF in the presence of missing data with testlets and the standalone items together.

Declarations

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

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

References

AERA, APA, and NCME (2014). *The standards for educational and psychological testing*. Washington, DC: American Psychological Association.

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

Allison, P. D. (2002). *Missing data*. Thousand Oaks, CA, US: Sage publications.

Banks, K. (2015). An introduction to missing data in the context of differential item functioning. *Practical Assessment, Research & Evaluation*, *20*(12), 1-10.

Banks, K., & Walker, C. (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, San Francisco, CA.

Beretvas, S. N., & Walker, C. M. (2012). Distinguishing differential testlet functioning from differential bundle functioning using the multilevel measurement model. *Educational and Psychological Measurement*, *72*(2), 200-223. https://doi.org/10.1177/0013164411412768

Chang, H.-H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *ETS Research Report Series*, *1995*(1), i:30.

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

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement*, *43*(2), 145-168. https://doi.org/10.1111/j.1745-3984.2006.00010.x

DeMars, C. E. (2012). Confirming testlet effects. *Applied Psychological Measurement*, *36*(2), 104-121. https://doi.org/10.1177/0146621612437403

Doğanay Erdoğan, B. (2012). *Assessing the performance of multiple imputation techniques for Rasch models with a simulation study*[Doctoral dissertation, Ankara University]. Council of Higher EducationThesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezDetay.jsp?id=RjkkLNSzxvPF7_el9Z6dkg&no=6IGSpCmQZHcSkxQALa295w

Douglas, J. A., Roussos, L. A., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement*, *33*(4), 465-484. https://doi.org/10.1111/j.1745-3984.1996.tb00502.x

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

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

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. (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

Fukuhara, H., &Kamata, A. (2011). A bifactor multidimensional item response theory model for differential item functioning analysis on testlet-based items. *Applied Psychological Measurement*, *35*(8), 604-622. https://doi.org/10.1177/0146621611428447

Garrett, P. L. (2009). *A monte carlo study investigating missing data, differential item functioning, and effect size*.) [Doctoral dissertation, Georgia State University]. https://scholarworks.gsu.edu/eps_diss/35/

Im, J., Cho, I. H., & Kim, J. K. (2018). *FHDI: Fractional Hot Deck and Fully Efficient Fractional Imputation*. https://cran.r-project.org/web/packages/FHDI/index.html

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*, 1030-1043.

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

Lee, Y.-S., Cohen, A., & Toro, M. (2009). Examining type I error and power for detection of differential item and testlet functioning. *Asia Pacific Education Review*, *10*(3), 365-375. https://link.springer.com/article/10.1007/s12564-009-9039-7

Ludlow, L. H., & O'leary, M. (1999). Scoring omitted and not-reached items: Practical data analysis implications. *Educational and Psychological Measurement*, *59*(4), 615-630. https://doi.org/10.1177/0013164499594004

Min, S., & He, L. (2020). Test fairness: Examining differential functioning of the reading comprehension section of the GSEEE in China. *Studies in Educational Evaluation*, *64*. https://doi.org/10.1016/j.stueduc.2019.100811

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

Ntzoufras, I. (2009). *Bayesian modeling using WinBUGS*. John Wiley & Sons.

Ravand, H. (2015). Assessing testlet effect, impact, differential testlet, and item functioning using cross-classified multilevel measurement modeling. *SAGE Open*, *5*(2). https://doi.org/10.1177/2158244015585607

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). *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, April 2004.

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

Sedivy, S. K. (2009). Using traditional methods to detect differential item functioning in testlet data. [Doctoral dissertation, University of Wisconsin-Milwaukee]. ProQuest Dissertations Publishing. https://www.proquest.com/openview/4ea81f321746d15a968b1505d7c8102b/1?pq-origsite=gscholar&cbl=18750

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 of Measurement in Education.

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

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). WinBUGS version 1.4.3 [Computer Program]. MRC Biostatistics Unit, Institute of Public Health.

Tamcı, P. (2018). *Investigation of the impact of techniques of handling missing data on differential item functioning*. [Master's Thesis, Hacettepe University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezDetay.jsp?id=TuYJvxte3q-jTBLlv22wLg&no=hOhq2SUN1BT96zGOfFKULw

Taşdelen Teker, G. (2014). *The effect of testlets on reliability and differential item functioning*. [Doctoral Dissertation, Hacettepe University]. Council of Higher Education Thesis Center.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*, 185-201. https://www.jstor.org/stable/1434630

Walker, C. M. (2011). What's the DIF? Why differential item functioning analyses are an important part of instrument development and validation. *Journal of Psychoeducational Assessment,29*(4), 364-376. https://doi.org/10.1177/0734282911406666

Wang, W.-c., & Wilson, M. (2005). Assessment of differential item functioning in testlet-based items using the Rasch testlet model. *Educational and Psychological Measurement*, *65*(4), 549-576. https://doi.org/10.1177/0013164404268677

Yang, S., & Kim, J. K. (2016). A note on multiple imputation for method of moments estimation. *Biometrika, 103*(1), 244-251. https://doi.org/10.1093/biomet/asv073

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*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.