International Journal of
Assessment Tools in Education

Volume: 5 Number: 1
January 2018

ISSN-e: 2148-7456 online Journal homepage: http://www.ijate.net/ http://dergipark.gov.tr/ijate

Investigating the Impact of Missing Data Handling Methods on the Detection of Differential Item Functioning

Hüseyin Selvi, Devrim Özdemir Alıcı

To cite this article: Selvi, H., Özdemir 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. DOI: 10.21449/ijate.330885

To link to this article: http://ijate.net/index.php/ijate/issue/view/13
http://dergipark.gov.tr/ijate

Int. J. Asst. Tools in Educ., Vol. 5, Issue 1, (2018) pp. 1-14

http://www.ijate.net
e-ISSN: 2148-7456 © IJATE
INTERNATIONAL JOURNAL OF ASSESSMENT TOOLS IN EDUCATION

Research Article

Investigating the Impact of Missing Data Handling Methods on the Detection of Differential Item Functioning

Hüseyin Selvi[*1], **Devrim Özdemir Alıcı**[*2]

[1]Mersin University, Medical Faculty, Medical Education Department, Turkey
[2]Mersin University, Faculty of Education, Department of Measurement and Evaluation in Education, Turkey

Abstract: In this study, it is aimed to investigate the impact of different missing data handling methods on the detection of Differential Item Functioning methods (Mantel Haenszel and Standardization methods based on Classical Test Theory and Likelihood Ratio Test method based on Item Response Theory). In this regard, on the data acquired from 1046 candidates who entered to Foreign National Student Exam (FNSE) held in year 2016 by Mersin University (MEU) and answered Basic Skills subtest, using different missing data handling methods, differential item functioning analyses with Mantel Haenszel, Standardization and Likelihood Ratio Test methods are performed. Basic Skills test consists of 80 multiple choice items. The items are all binary scored (1-0) items. Among the participants 523 are female and 523 are male. The findings showed that the number of items flagged as DIF has changed with the used missing data handling methods. The DIF detection methods based on Classical Test Theory are more consistent within themselves compared to DIF detection method based on Item Response Theory, whereas the used missing data handling methods differentiate the DIF detected items and this difference reaches a significant level for Mantel Haenszel method

ARTICLE HISTORY

Received: 31 March 2017
Revised: 30 June 2017
Accepted: 23 July 2017

KEYWORDS

Differential Item Functioning DIF; Test and Item Bias, Missing Values; Imputation of Missing Data; Mantel Haenszel; Likelihood Ratio Test

1. INTRODUCTION

Even if the reliability of the measurements acquired with a measurement tool is investigated with different method, in some cases where the desired quality (latent trait) to be measured is mixed with other qualities, the individuals in different subgroups can be affected systematically from this situation. In the current literature it is named as “bias” and causes negative effect on validity due to the definition, and it decreases somehow the reliability.

This study was presented as an oral presentation at 2016 international 5. Measurement and Evaluation Conference at Antalya.

*Corresponding Author E-mail: hsyn_selvi@yahoo.com.tr

ISSN: 2148-7456 online /© 2018 **DOI: 10.21449/ijate.330885**

Bias that occurs as a systematic variation source and affects the validity is defined as "the difference between the probabilities of correct answer of the individual within different subgroups with the same ability level (Angoff, 1993).

From this definition, in the studies regarding the determination of the bias initially, it is understood that it is necessary to match the individuals in different subgroups regarding the ability levels and to examine statistically the item parameters of these individuals. This situation is defined as the examination of whether there is Differential Item Function (DIF) in the items or not.

It is required that the items with detected DIF should be checked by the experts and whether the DIF is due to another source rather than the desired measured quality shall be investigated. In cases that the DIF is detected to be caused by another source than the desired measured quality, it can be convinced of that the related item(s) is/are biased (Camilli & Shepard, 1994; Zumbo, 1999).

In order to provide validity of the items detected biased, it can be said that it is proper for them to be revised in possible cases, and in impossible cases to be removed completely from the test. In fact, in the literature it is described that one of the important threats that affect the objectivity and validity of the measurement tools is the bias (Kristanjansonn, Aylesworth, McDowell & Zumbo, 2005).

Bias, besides decreases the validity, presents a preventable structure as a systematic variation source. Thus, scientists have developed significantly extensive methods regarding the detection of DIF. As examples of some frequently used ones of these methods Standardization (SPD-X), Mantel-Haenszel (M-H), Logistic Regression (LR) and Likelihood Ratio Test (LRT) methods can be given (Angoff, 1993; Camilli&Shepard, 1994; Osterlind, 1983).

However, it is possible to say that nearly all of these frequently used methods and other methods have different weaknesses and strengths and many methods are developed to fix weakness of each. Hence, in DIF detection there are many different distresses like in methods acting over item difficulty (pj) index, 'pj' values are affected from the average group differences and item discrimination index (rjx). In methods based on variance analysis, variance to be affected from pj and rjx values, in methods based on correlation, 'rjx' is able to be able to process in similar ways for the groups and even if the 'pj' differs, in this case to increase correlation coefficient, the correct response likelihood of the item to operate in favor of the same group for all ability levels and non-uniform DIF situation to arise etc. (Selvi, 2013).

In addition to these in the literature, studies are showing the different DIF detection methods also being affected from many variables like number-ratio of items with DIF, test length, DIF level, sample size, DIF structure in items, and item scoring method etc. (Camili & Shepard, 1994; Gelin & Zumbo, 2003; Gierl, Jodoin & Ackerman, 2000; Narayanan & Swaminathan, 1994; Osterlind, 1983; Padilla, Hidalgo, Benitez & Gomez-Benito, 2012; Selvi, 2013).

Another variable that can change the findings acquired by the DIF detection methods is thought to be the problem of missing data. Hence, many statistical methods used today based on complete data matrix and missing data rate being increased may cause these methods to give erroneous results (Bernhard, Celia & Caotes, 1998; Molenberghs & Kenward, 2007; Woodward, Smith & Tunsatall-Pedoe, 1991).

Similarly, in the literature, including M-H, LR, SIBTEST, it is said that many DIF detection methods are not capable of handling missing data (Banks, 2015). Missing data can be formed in cases like, for a performance test not reaching the item due to time limitations, accidentally

omitting the item or leaving it empty due to not knowing the right answer (Banks, 2015); for a scale, accidentally omitting the related item or refusal to answer due to personal reasons. In other words, and in the most general sense, the missing data can be considered as an information loss (Alpar, 2011).

Missing data may lead to problems like decrease of the power of the used statistical analyses, faulty estimate of standard error, increase in Type I error rate, not being able to estimate in quality the closed properties based on observation (Hohensinn & Kubinger, 2011; Molenberghs & Kenward, 2007). Thus, many studies have been done in line with the resolution of the missing data problem in time and many different methods have been developed.

Regarding the proper method to be chosen, primarily the pattern and the mechanism of the missing data should be understood. For this aim the issues like whether the missing data is distributed over the observations randomly, whether they have a specific pattern, how much missing data there is (how frequently it occurs) etc. are investigated. In other words, it is researched whether there is a case leading to missing data process in the data or not is researched (Alpar, 2011). In the literature regarding this process, it is mentioned that researchers acting carefully in data collecting presents an opportunity in observing the reasons and increasing the quality of the possible missing data (Pigott, 2001).

On the other hand, the researchers in general act in tendency to prove the assumption that the missing data does not make a significant difference on the study findings and can perform listwise deletion of the missing data with the assumption that it is missing at random (MAR) without investigating whether it is negligible or not (Alison, 2002; Groves, 2006).

In ignoring the missing data problem (un)consciously, it is thought that conditions like the researcher not having sufficient knowledge on the field of missing data problem, in scoring of the measuring tools where the maximum performance are measured (especially in optic reader usage) 1 point to correct answered items and 0 points to be assigned to the incorrect, left empty or different marking done items thus the missing data being removed by zero imputation method somehow without examination, in some statistical software the missing data to be removed by a default method automatically etc. are in play. This condition is specially emphasized in a study done by Demir & Parlak (2012). In the related study 405 researches conducted in Turkey universe and containing statistical analysis process are examined and in 40% of these studies, despite containing different analysis methods like standard error, mean, variance, covariance, correlation, t and F statistic, reliability and validity coefficients, factor analysis, regression analysis, structural equation modelling analyses, it is indicated that there was no explanation/proof seen regarding whether the data set on which the analyses are conducted had missing data or not. Listwise deletion and zero imputation make the resolution of the problem fairly ease in cases that the missing data is really formed as missing at random. However, any method to be used before the quality of the missing data is understood also consists of the possibility that the study findings are faulty.

Rubin (1976) defined three possible conditions regarding the understanding of the quality of the missing data (Missing Data Mechanism). These define cases in which the missing data is formed as missing completely random, MCAR, missing at random, MAR, and missing at non-random, MNAR. MCAR explains the situations that the probability of a value regarding x variable to be a missing data is not related to x variable itself or any value regarding another variable in the data set (Alison, 2002). In other words, MCAR explains the cases where there are no justified explanations is made regarding the formation of the missing data and the formation of the missing

data is referenced to randomization (Peng & Zhu, 2008). When the condition is looked at from DIF angle, Banks (2015) says that the MCAR missing data formation is realized in general when the related item is left empty both by the focus and the reference groups accidently.

MAR expresses the cases where the probability of a value regarding x variable to be a missing data is not related to x variable itself when the other variables in the data set are fixed (Alison, 2002). In other words, MAR is the cases in which the probability of missing data formation in the certain item is related to the observed data systematically. In the perspective of DIF definition, this situation is explained as for a test includes 30 items, the probability of the DIF analyzed items without any response (empty items) is dependent on which group that the individuals are in (focus, reference) or their performances in 2nd - 29th items (Peng & Zhu, 2008).

MNAR is the cases where the probability of a value regarding x variable to be a missing data is related to x variable itself. In other words, MNAR explains the cases where the probability of individuals to leave the item empty depends on the performances of individuals on the related item, item being left empty as it is faulty etc. (Peng & Zhu, 2008). Alison (2002), based on the definitions Rubin (1976) made regarding the quality of the missing data, classified the missing data simply as ignorable and nonignorable. In order for the missing data to be ignored, Alison said that it should be in MAR or MCAR and a missing data in MNAR cannot be ignored. Here, by the ignorable term means the case where extra modelling of missing data is not needed for the analyses to be made.

In the literature search regarding the missing data problem, there are many studies suggesting a resolution of this problem and many different methods have been developed. These methods in general are classified within as methods based on deletion and value assignment (Alpar, 2011; Demir, 2013; Alison, 2002; Little & Rubin, 1987). Among methods based on deletion; listwise deletion and analysis wise deletion, among methods based on value assignment (simple); zero imputation, mean substitution, assigning mean of nearby points, assigning median of nearby points and regression imputation methods are used frequently in the literature (Banks, 2015; Little & Rupin, 1987; Alison, 2002; Alpar, 2011).

In *listwise deletion method*; the observations containing one or more missing data are removed from the data.

In *analysiswise deletion method*; observation(s) or variables with missing data are removed from the analysis if only they are to be analyzed.

As seen, deletion methods appear as fairly simple approaches regarding the resolution of the missing data problem. However, removing the missing data from the observation via deletion methods can cause serious decrease in observation numbers and a sample deemed sufficient can turn into a sample with insufficient numbers. Moreover, methods based on deletion can decrease the stability of the calculated statistics, can place the validity and generalizability of the study to distress (Alpar, 2011). In addition to this, for methods based on deletion to be used, the assumption of missing data being in MCAR should be met (Alison, 2002; Alpar, 2011).

In *methods based on value assignment*, new values are assigned to the missing values based on specific assumptions and rules. In assigning these values (except zero imputation method) the other values or variables in the data set are considered.

In *zero imputation method*, omitted item is considered as ‘wrong’ or in most general state ‘zero’ points are assigned to this value. However, as this condition leads to biased parameter estimates and faulty hypothesis results, in Item Response Theory (IRT) and DIF studies it is especially not recommended (Banks & Walker, 2006; Lord, 1974).

In *mean imputation method*, empty value(s) is/are filled via taking the average of the values given by other individuals to the related item as serial mean imputation, via taking the average of the values given to other items by the individual as unit's mean imputation, via taking mean of nearby points, via taking median of nearby points etc. However, this condition too, can cause bias addition to many analysis results including variance-covariance estimates and parameter estimates (Little & Rupin, 1987). Similarly, for these assignment methods to be used the assumption of missing data being in MCAR should be met (Alpar, 2011).

The regression imputation method; is based on estimation operations realized by taking the regressed variable as the variable with missing value(s) and other variable(s) as regressing variables. However, in this method, as it starts upon relations between other variables, the already present relation in the data can be strengthened more as the result of the assignment thus lead to being biased. In addition, the value obtained as the result of estimation can exceed the score range of the missing data. In order to use the regression imputation methods, the missing data being in MCAR should again be met (Alpar, 2011).

The methods based on deletion and value assignment appear as frequently used method in resolution of the missing data problem. However, it is known that these methods also bring up many restrictions. These restrictions, whereas, drove the researchers to develop new methods.

Among the methods suggested in this regard, the multiple imputation method suggesting estimation of the missing data via using two or more methods together and Expected-Maximization method based on maximum likelihood shine out are mentioned (Alison, 2002; Alpar, 2011; Demir, 2013; Little & Rubin, 1987). The most important advantage of these methods compared to methods based on deletion and simple value assignment is that they can also be used in cases where the missing data is in MAR (Alison, 2002; Alpar, 2011).

When the studies performed in literature regarding the missing data problem and used methods are examined; it is suggested that in cases that may cause serious reduction in data set or bias listwise deletion shall not be used (Graham, 2009). As it increases Type I error rate the zero imputation method shall be avoided if possible (Banks & Walker, 2006; Banks, 2015; Robitzsch & Rupp, 2009). The method with Type I error rate that is similar to the complete data set shall be preferred (Banks & Walker, 2006; Finch, 2011) and especially in DIF studies the missing data problem shall not be ignored (Banks, 2015). Besides; it is expressed that sample size and DIF level in items being increased, the performance of analysiswise deletion methods instead of listwise deletion and zero imputation methods, increase the rate of accurately determined items with DIF. It is shown that item to grow difficult and missing data rate to be increased decreases as well (Banks & Walker, 2006; Emenogu, Falenchuck & Childs, 2010; Finch, 2011; Garrett, 2009).

On the other hand, the most efficient solution in missing data problem can be shown as with precautions like being careful yet on the data gathering stage, training individual given the task of data gathering, the missing data not to be present or be in ignorable quality and level (Alison, 2002; Little & Rubin, 1987). In this regard, there are different suggestions in literature regarding the ignorable missing data ratio. Schafer (1999) said that this rate should be below 5%, Bennett (2001) 10%, Peng, Harwell, Liou & Ehman (2006) 20% and otherwise it should be considered that the findings acquired from the study may be biased.

The missing data problem and DIF are still seen important problem and research studies on these topics are ongoing. In the literature there are many extensive studies regarding the detection of the lacking and powerful points of the missing data approaches and DIF detection methods.

However, it is observed that nearly all of these studies were performed over data sets acquired by simulation method (e.g., Banks & Walker, 2006; Banks, 2015; Emenogu, Falenchuck & Childs, 2010; Falenchuck & Herbert, 2009; Finch, 2011; Garrett, 2009; Hohensinn & Kubinger, 2011; Pigott, 2001; Robitzsch & Rupp, 2009; Rousseau, Bertrand & Boiteau, 2006; Sedivy, Zhang & Traxel, 2006). And it is observed that nearly all of these studies were performed over frequently used DIF detection methods like, Standardization, SIBTEST, Linear Logistic Regression and Likelihood Ratio Test (e.g., Banks, 2015; Finch, 2011; Robitzsch & Rupp, 2009; Wu, Lee & Zumbo, 2007). A study which includes the Classical Test Theory (CTT) and Item Response Theory (IRT) based DIF detection methods, a non-simulative data set and expected maximization and regression imputation methods at the same time is not seen.

In the literature, regarding the studies conducted on simulation technique, it is expressed that being aware of the situation that these studies cannot present enough proof that the actual results shall be found and cannot guarantee the accuracy of the results to be found and thus it is imperative to be sure exactly that all the analytic and experimental options that can be used in solving the problem would not be usable before these studies are performed and finally they should be used as last resort (Harwell, Stone, Hsu & Kirisci, 1996).

Thus in this study, the answer of the question "How are the performances of expectation maximization and regression imputation methods for handling with missing data on detecting DIF methods based on CTT and IRT is sought.

2. METHOD

In this study, over the complete data matrix obtained by using different missing data methods, the investigation of operation of DIF detection methods based on different theories in regard to gender variable is aimed for. Thus it can be said that the type of this study is basic research (Kothari, 2004; Royce, Straits & Straits, 1993; Singh, 2006).

Data acquired from 1046 candidates who attended to the Foreign National Student Exam (FNSE) conducted by Mersin University (MEU) in year 2016 and answered Basic Learning Skills subtest.

Some descriptive information related to the participants is given in Table 1.

Table 1. Descriptive values regarding the participating group

	Foreign National	Turkish National	Total
Female	448 (50.1%)	75 (49.3%)	523 (50%)
Male	446 (49.9%)	77 (50.7%)	523 (50%)
Total	894 (100%)	152 (100%)	1046 (100%)

2.1. Instrument

FNSE consists of two subtests as Basic Skills Test and a Language Test and is applied to high school graduates in Turkey and specific centers around the world every year for granting them undergraduate education in MEU. Candidates are ranked according to the scores they achieved in this exam and regarding specific quotas, are placed to programs they chose. In

development of the tests, all works are planned and realized by the Measurement and Evaluation Application and Research Centre of the university. The Basic Skills subtest was used as data collecting tool in this study is scored in binary (0-1), multiple-choice and consists of 80 items with 5 choices and the reliability (KR 20) of the acquired scores is calculated as 0.95.

2.2. Data Analysis

DIF analyses was done via M-H, Standardization and LRT methods. M-H, and Standardization methods do not contain the assumptions which of the parametric techniques should be faced. However, as LRT is one of the methods based on IRT the data should meet the unidimensionality and local independence that are basic assumptions of IRT (Embretson & Reise, 2000; Hambletton & Swaminathan, 1985). Thus, in the first stage of the data analysis whether these assumptions were checked.

In this regard, the unidimensionality that is one of the basic assumption of the IRT, is investigated utilizing the principal components analysis based on intra-item tetrachoric correlation matrix and the data is observed to be unidimensional from the acquired results regarding the local independence, in the literature it is said that this assumption is linked to the unidimensionality and a data that is seen to be unidimensional meets also the local independence (Lord, 1980: 19; Hambletton & Swaminathan, 1985: 25). Based on these it is deemed that the study data also meets the local independence.

In the second stage of the data analysis, in order the analysis based on Item Response Theory to be done, model-data fit was examined. Because the likelihood ratio test, which is one of the DIF methods used in this research based on IRT and the DIF analysis software (IRT-LR-DIF) requires the selection of the model. The -2 log likelihood value of the data obtained for the two parameter logistic model is calculated as 71207,78. As this statistic showing χ^2 distribution is very sensitive to sample size and in big sample sizes model-data fit cannot be provided for nearly all models; for evaluation of the model data fit $-2 \log likelihood / (S-1) - 2n(r-1) \leq 3.00$ condition is considered. Here '*S'* shows response pattern number, *n* number of items, *r* number of response category The possible response pattern of this study dependent on the item number and response category number is 5^{80}. Bock (1997) indicates that all values meeting the '–2 log likelihood / (S-1) – 2n(r-1) ≤ 3.00' condition are sufficient for model data fit (Gözen Çıtak, 2007). Based on these findings it can be said that the data is fit to the 2 parameter logistic model.

In the third stage of the data analysis, in order to decide the pattern of the missing data Little's MCAR test was applied and it was observed that the data was not in MCAR ($\chi 2$=22815.65, p<0.05). In the fourth stage of data analysis, the missing values that are present in the raw data set and whose ratios change in between 0.3% and 10%, due to the data not being in MCAR, are removed by Expectation Maximization and Regression Imputation and DIF analyses are made on complete data set by Mantel Haenszel, Standardization and Likelihood Ratio Test methods and items showing DIF and number of items with DIF are determined.

Whether the number of items determined with different missing data methods and different DIF detection methods show discrepancies is examined by Cohran's Q and McNemar tests. Cohran's Q test is used for testing whether the number of items with DIF determined via Mantel Haenszel, Standardization and Likelihood-Ratio Test for each missing data method, differentiate from each other or not; and McNemar test is used if there is a significant difference found by

Cohran's Q test and in order to test whether the number of DIF included items according to the used missing data method are significantly different from each other or not.

3. FINDINGS

In the scope of the study the DIF analyses performed on the complete data matrix obtained by expectation maximization and regression imputation methods and the values obtained as the result of these analyses are given in Table 2.

Table 2. Results of DIF analysis performed on complete data matrix obtained by expectation maximization and regression imputation methods.

Items	Expectation Maximization						Regression Imputation						Missing Data Ratio (%)
	Focus-Ref. Group Mean		M-H		Std.	LRT	Focus-Ref. Group Mean		M-H		Std.	LRT	
	Male	Female	$MH\chi^2$	*p*	*SPD**	G^2**	Male	Female	$MH\chi^2$	*p*	*SPD**	G^2**	
Item 1	0.9	0.87	3.67	0.05	-0.03	5.5	0.9	0.87	5.21	**0.02**	-0.03	**6.6**	1.0
Item 2	0.87	0.87	0.00	0.92	-0.01	0.1	0.86	0.87	0.09	0.75	0.00	0.10	1.3
Item 3	0.75	0.73	0.27	0.59	0.00	1.3	0.75	0.73	0.47	0.49	-0.01	1.6	1.9
Item 4	0.84	0.78	6.72	**0.01**	**-0.05**	**11.8**	0.83	0.78	7.01	**0.00**	**-0.06**	**12.2**	2.4
Item 5	0.85	0.85	0.05	0.81	0.00	1.4	0.85	0.84	0.1	0.74	0.00	1.6	1.7
Item 6	0.69	0.69	0	0.92	0	1.6	0.68	0.69	0.13	0.71	0.01	1.5	3.3
Item 7	0.44	0.41	0.39	0.53	-0.02	1.9	0.45	0.42	0.25	0.61	-0.02	2	6.8
Item 8	0.93	0.92	0	0.95	0	0.9	0.93	0.92	0.01	0.91	0	1.1	0.4
Item 9	0.62	0.63	0.02	0.87	0	0.4	0.62	0.63	0.15	0.69	0.01	0.1	7.8
Item 10	0.88	0.91	1.99	0.15	0.02	**6.9**	0.88	0.91	3.2	0.07	0.04	5.4	1.1
Item 11	0.66	0.55	19.81	**0**	**-0.12**	**22.8**	0.64	0.54	14.3	**0**	**-0.10**	**17.9**	5.3
Item 12	0.54	0.5	1.76	0.18	-0.04	**6.6**	0.54	0.51	1.12	0.28	-0.02	5.2	3.5
Item 13	0.81	0.85	1.15	0.28	0.02	3.4	0.81	0.85	1.22	0.26	0.02	4.5	1.6
Item 14	0.55	0.53	0.45	0.49	-0.02	1.8	0.56	0.54	0.46	0.49	-0.01	1.5	5.3
Item 15	0.91	0.89	0.17	0.67	-0.01	1.8	0.9	0.89	0.22	0.63	0	2	0.8
Item 16	0.82	0.86	2.98	0.08	0.02	3.7	0.82	0.86	3.22	0.07	0.03	4	1.1
Item 17	0.93	0.94	1.37	0.24	0.01	2.6	0.93	0.94	0.42	0.51	0	1.7	0.4
Item 18	0.9	0.89	0.11	0.73	0	0.8	0.9	0.89	0.07	0.78	-0.01	0.5	0.7
Item 19	0.91	0.92	0.67	0.41	0.01	2.8	0.92	0.93	0.23	0.62	0.02	2.6	0.3
Item 20	0.93	0.95	4.39	**0.03**	0.04	2.6	0.93	0.94	2.04	0.15	0.02	2.2	0.5
Item 21	0.27	0.25	0.3	0.58	-0.02	0.7	0.28	0.25	0.32	0.56	-0.02	0.2	2.2
Item 22	0.75	0.76	0.09	0.75	0.01	0	0.75	0.75	0.01	0.90	0	0	3.0
Item 23	0.66	0.63	2.1	0.14	-0.03	5.3	0.64	0.62	1.85	0.17	-0.03	3.5	6.8
Item 24	0.82	0.76	5.94	**0.01**	**-0.06**	**8.3**	0.82	0.76	5.99	**0.01**	**-0.06**	**7.1**	1.2
Item 25	0.87	0.88	0.27	0.6	0	0.3	0.87	0.88	0.56	0.45	0.02	0.8	2.2
Item 26	0.69	0.63	3.99	**0.04**	**-0.06**	**8.7**	0.69	0.62	7.85	**0**	**-0.07**	**11.3**	3.7
Item 27	0.87	0.86	0.53	0.56	0	**6.8**	0.87	0.86	0.88	0.34	-0.01	**6.4**	0.9
Item 28	0.92	0.93	0.02	0.88	0	0.5	0.92	0.93	0.14	0.70	0.01	0.5	0.9
Item 29	0.9	0.9	0	0.93	0	1.9	0.9	0.9	0.48	0.48	-0.01	2.1	1.3
Item 30	0.66	0.66	0.02	0.88	0	0.9	0.66	0.64	1.44	0.22	-0.03	1.8	5.9
Item 31	0.84	0.84	0	0.98	-0.01	0.7	0.84	0.84	0	0.93	0	1.5	1.9
Item 32	0.53	0.47	3.66	0.05	**-0.05**	**10.3**	0.53	0.46	3.91	**0.04**	**-0.05**	**10.9**	5.6
Item 33	0.69	0.72	0.73	0.39	0.03	0	0.69	0.71	0.95	0.32	0.03	0	4.3
Item 34	0.8	0.83	0.7	0.40	0.02	1	0.81	0.83	0.04	0.83	0.01	0.2	1.8
Item 35	0.64	0.65	0	0.98	0	4.8	0.64	0.64	0	0.95	0	4.6	5.0
Item 36	0.33	0.31	0.02	0.88	0	2.2	0.65	0.33	0.22	0.63	0	3.3	5.7
Item 37	0.69	0.74	3.25	0.07	**0.05**	2.9	0.69	0.73	1.82	0.17	0.03	2	5.3
Item 38	0.93	0.92	0.02	0.88	0	5.7	0.93	0.92	0.12	0.72	-0.01	**6.6**	1.0

Item 39	0.63	0.63	0.04	0.83	0	0.3	0.63	0.63	0.2	0.64	-0.01	0	3.2
Item 40	0.9	0.89	0.15	0.69	0	2.1	0.9	0.89	0.13	0.71	0	2.2	0.4
Item 41	0.7	0.7	0	0.95	0	0	0.7	0.7	0	0.95	0	0	1.2
Item 42	0.88	0.9	0.56	0.45	0.02	1.5	0.88	0.9	0.76	0.38	0.01	1.5	0.8
Item 43	0.75	0.75	0.04	0.82	0	0.7	0.75	0.75	0	0.95	0	0.6	1.8
Item 44	0.8	0.73	10.78	**0**	**-0.08**	**11.7**	0.8	0.74	9.64	**0**	**-0.07**	**8.9**	1.6
Item 45	0.47	0.42	3.73	0.05	**-0.05**	**9.9**	0.48	0.43	4.1	**0.03**	**-0.05**	**8.1**	4.4
Item 46	0.16	0.2	1.62	0.20	0.03	1.5	0.16	0.2	1.46	0.22	0.03	0.5	2.3
Item 47	0.82	0.84	0.35	0.55	0	0	0.82	0.84	0.12	0.72	0	0.1	2.4
Item 48	0.74	0.76	0.19	0.65	0.01	1.6	0.73	0.76	0.07	0.78	0	1.8	3.0
Item 49	0.71	0.78	8.08	**0**	**0.05**	**8.7**	0.71	0.79	8.61	**0**	**0.05**	**11.1**	3.1
Item 50	0.3	0.63	1.08	0.29	0.02	1.2	0.59	0.62	0.25	0.61	0.01	0.7	8.3
Item 51	0.7	0.76	2.39	0.12	0.04	4.8	0.7	0.76	4.91	**0.02**	**0.05**	5	4.7
Item 52	0.76	0.82	8.27	**0**	**0.05**	**8.4**	0.76	0.82	4.14	**0.04**	0.03	**6.2**	4.1
Item 53	0.69	0.7	0.06	0.8	-0.01	1.6	0.69	0.69	0.32	0.57	-0.01	3.6	4.8
Item 54	0.8	0.88	10.06	**0**	**0.05**	**19.1**	0.8	0.87	10.1	**0**	**0.05**	**17.4**	2.3
Item 55	0.63	0.62	1.35	0.24	-0.02	4.5	0.63	0.62	0.52	0.46	-0.01	3.7	6.6
Item 56	0.67	0.71	1.43	0.23	0.02	3.4	0.67	0.71	2.02	0.15	0.03	3.2	4.5
Item 57	0.56	0.61	1.76	0.18	0.03	2.9	0.56	0.6	1.63	0.20	0.03	2.9	6.6
Item 58	0.72	0.74	0.08	0.77	0	2.5	0.72	0.74	0.02	0.87	0	1.5	3.8
Item 59	0.55	0.61	3.78	0.05	0.04	4	0.54	0.6	4.74	**0.02**	**0.06**	5.7	6.9
Item 60	0.39	0.25	19.19	**0**	**-0.12**	**28.7**	0.39	0.26	15.8	**0**	**-0.11**	**25.9**	4.2
Item 61	0.69	0.73	2.7	0.09	0.03	2.1	0.69	0.72	1.18	0.27	0.02	2.6	5.1
Item 62	0.64	0.64	0.1	0.74	-0.02	3.7	0.63	0.63	0.07	0.78	-0.02	2.8	6.0
Item 63	0.69	0.74	3.61	0.05	0.03	2.5	0.68	0.74	4.08	**0.04**	0.03	2.9	5.2
Item 64	0.45	0.43	0.12	0.72	-0.01	4.5	0.47	0.45	0.14	0.70	-0.01	3.1	9.1
Item 65	0.76	0.76	0	0.94	0	3.2	0.76	0.76	0.01	0.89	0	1	4.8
Item 66	0.35	0.34	0.07	0.78	-0.01	0.4	0.37	0.36	0.16	0.68	-0.01	0.3	3.7
Item 67	0.15	0.17	0.04	0.83	0	3.7	0.17	0.18	0.11	0.73	0	4	10.3
Item 68	0.47	0.49	0.64	0.42	0.03	0.7	0.47	0.49	1.46	0.22	0.04	1.1	7.8
Item 69	0.49	0.5	0.07	0.78	0	5.2	0.49	0.51	0.67	0.41	0.01	1.5	9.0
Item 70	0.58	0.61	1.86	0.17	0.02	0.8	0.58	0.61	0.33	0.56	0	0.1	7.5
Item 71	0.54	0.57	0.73	0.38	0.02	1	0.54	0.57	1.17	0.27	0.03	3	8.9
Item 72	0.63	0.64	0	0.97	-0.01	**12.7**	0.64	0.64	0.18	0.66	-0.01	**15**	6.1
Item 73	0.54	0.56	0.16	0.68	0	5.2	0.54	0.56	0.19	0.65	0	3.7	8.4
Item 74	0.62	0.62	0.26	0.60	-0.01	2.6	0.63	0.63	0	0.92	0	1.9	5.5
Item 75	0.32	0.34	0.91	0.33	0.02	0.5	0.33	0.37	3.31	0.06	0.04	3.3	7.9
Item 76	0.65	0.71	2.36	0.12	0.04	2.6	0.65	0.71	3.64	0.05	**0.05**	2.6	5.8
Item 77	0.54	0.54	0.07	0.77	0	0.5	0.55	0.54	0	0.92	-0.01	0.9	7.3
Item 78	0.4	0.37	1.28	0.25	-0.02	4.6	0.42	0.39	0.35	0.54	-0.02	3.4	10.0
Item 79	0.69	0.77	6.19	**0.01**	**0.05**	**11.1**	0.68	0.75	7.27	**0**	**0.06**	**9.6**	5.8
Item 80	0.58	0.59	0.05	0.81	0.01	0	0.56	0.58	0.5	0.47	0.02	0.5	7.2

***** SPD-X values are located between '-1.00' to '1.00'. The values between -0.05 to 0.05 shows ignorable level of DIF; and values between -1 to -0.05 and 0.05 to 1 intervals shows unignorable level of DIF presence (Gonzales. Padilla. Dolores. Gomez & Benitez. 2010).

**As the G^2 values calculated with LRT test show the chi-square distribution in the freedom degree up to estimated parameter number. the critical value of the chi-square distribution here regarding the DIF detection is taken as 5.99 (p=0.05. df=2) (Dişçi. 2012).

When Table 2 is examined. in the analyses performed on the complete data matrix obtained by expectation maximization. DIF is seen in 11 items with M-H method. 13 items with Standardization method and 16 items with LRT method. Similarly, on the complete data matrix obtained by regression imputation method. DIF is seen in 16 items with M-H method. 14 items with Standardization method and 16 items with LRT method.

The results of Cohran's Q and McNemar tests performed regarding whether the items determined with different missing data methods and different DIF detection methods show difference and simple coefficient of concordance calculated related to these are shown in Table 3 and Table 4.

Table 3: The results of Cohran's Q and McNemar tests performed regarding whether the items determined with different missing data methods and different DIF detection methods show difference.

Missing Data Methods	MH. Std. and LRT		MH (em-reg.)	Std. (em-reg.)	LRT (em-reg.)
	Cohran's Q	p	McNemar (p)	McNemar (p)	McNemar (p)
Expectation Max.	4.75	0.09	0.03	1.00	0.34
Regression Imputation	0.89	0.64			

Table 4. Simple coefficient of concordance calculated related to items determined with different missing data methods and different DIF detection methods

	MH. Std. and LRT	MH-Std.	MH-Std.	Std.-LRT
Expectation Max.	0.91	0.95	0.91	0.95
Regression Imputation	0.91	0.95	0.93	0.91

When Table 3 is examined. in the analyses performed on the complete data matrix obtained by both expectation maximization and regression imputation according to the Cochran's Q test results. items determined to be with DIF are observed to be differentiated from each other significantly. McNemar's test results show that the items determined by M-H method are differentiated significantly with the used missing data method. In the other DIF detection methods examined in the scope of the study in items with DIF determined regarding the used missing data method there has no significant change occurred.

The findings acquired in the scope of the study showed that the item numbers showing the DIF are changed among the DIF detection method. the DIF detection methods that are used in the scope of the study and based on the Classical Test Theory are more fit internally compared to the DIF detection method based on IRT. the used missing data approaches differentiate the items determined to be with DIF and this difference reaches to a significant level for Mantel Haenszel method.

4. DISCUSSION, CONCLUSION AND SUGGESTIONS

The findings acquired in this study showed that the items included DIF and their numbers were changed based on DIF detection method. The findings are partially overlapping with the findings of the other studies in the literature (Abedlazeez, 2010; Doğan & Öğretmen, 2008; Finch, 2011; Hohensinn & Kubinger, 2011; Kan, Sünbül & Ömür, 2013; Pigott, 2001; Robitzsch & Rupp, 2009; Spray & Miller, 1994; Ward & Bennett, 2012). Hence, in many of these studies significant difference between the items determined with different DIF methods and their numbers are present, whereas the determined difference in this study did not reach a significant level. Among the reasons, the difference between the item difficulty values obtained from the focus and reference

groups to be very close to zero, the related items and the test to be possibly qualified as ‘easy’ by the item difficulty value averages can be shown.

On the other hand, even if there was no significant difference between the results of DIF methods used for the missing data methods, the methods based on CTT are observed to have more concordance within compared to the methods provided by the methods based on IRT. The main reason of this can be shown as the M-H and Standardization methods to be calculated over contingency table and based on the same theory. These findings are overlapping with the findings of Selvi (2013).

In addition to these it is seen from the acquired findings that the used missing data approaches differentiate the items determined to be with DIF and this difference reaches to a significant level for Mantel Haenszel method. The findings acquired are overlapping with the findings of Robitzsch and Rupp (2009). In short, based on the findings obtained in the scope of this study and related literature, the conclusion can be reached that the used missing data approach, being also dependent on the DIF detection method, differentiate/can differentiate the items determined to be with DIF.

This result shows the possibility of the findings to be erroneous of the studies in which the missing data pattern and mechanism are ignored consciously/unconsciously or an inappropriate missing data approach is chosen and this reduces the importance of the missing data problem to an extent. The findings obtained in the scope of this study are limited with the expectation maximization and regression imputation methods among missing value assignment methods; and Mantel Haenzsel, Standardization and Likelihood Ratio Test methods among the DIF detection methods. Thus it can be suggested that similar studies, considering also the variables like scoring condition, sample size, different psychometric properties of items etc., shall be repeated with different missing data assignment method. Different DIF detection methods and the operation of different missing data methods on DIF shall be examined in order to contribute in solution of the missing data problem**.**

5. REFERENCES

Abedlazeez, N. (2010). Exploring DIF: Comparison of CTT and IRT methods. *International Journal of Sustainable Development*, 7(1), 11-46.

Allison, P. D. (2002). *Missing data.* California: Sage Publication Inc.

Alpar, R. (2011). *Uygulamalı çok değişkenli istatistiksel yöntemler.* Ankara: Detay Yayıncılık.

Angoff, W.H. (1993). Perspectives on differential item functioning methodology. In Holland & Wainer (Ed.), *Differential Item Functioning*. New Jersey: Lawrence Erlbaum Associates Publishers.

Banks, K., & Walker, C. (2006). *Performance of SIBTEST when focal group examinees have missing data.* San Francisco: National Council of Measurement in Education.

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

Bennett, D. A. (2001). How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health*, *25*, 464–469.

Bernhard, J., Celia, D.F., &Coates, A.S. (1998). Missing quality of life data in cancer clinical trials: Serious problems and challenges. *Statistics in Medicine*, 17, 517-532.

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

Demir, E., & Parlak, B. (2012). Türkiye'de eğitim araştırmalarında kayıp veri sorunu. *Journal of Measurement and Evaluation in Education and Psychology 3*(1), 230-241.

Demir, E. (2013). Kayıp verilerin varlığında çoktan seçmeli testlerde madde ve test parametrelerinin kestirilmesi: SBS örneği [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.

Dişçi, R. (2012). *Temel ve klinik biyoistatistik.* İstanbul: Tıp Kitapevi.

Doğan, N., & Öğretmen, T. (2008). Değişen Madde Fonksiyonunu belirlemede Mantel–Haenszel, Ki-Kare ve Lojistik Regresyon tekniklerinin karşılaştırılması. *Education and Science*, 33(148).

Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists*. London: Lawrence Erlbaum Associates.

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.

Falenchuk, O., & Herbert, M. (2009). *Investigation of differential non-response as a factor affecting the results of Mantel-Haenszel DIF detection* California: American Educational Research Association.

Finch, W.H. (2011). The impact of missing data on the detection of nonuniform differential ıtem functioning. *Educational and Psychological Measurement, 71*(4) 663–683.

Garrett, P. L. (2009). *A monte carlo study investigating missing data, differential item functioning, and effect size*. Georgia State University, Unpublished doctoral dissertation.

Gelin, M.N. & Zumbo, B.D. (2003). Differential item functioning results may change depending on how an item is scored: an illustration with the center for epidemiologic studies depression scale. *Educational and Psychological Measurement, X(X)* DOI: 10.1177/0013164402239317.

Gierl, M.J., Jodoin, M.G., & Ackerman, T.A. (2000). *Performance of Mantel-Haenszel, Simultaneous Item Bias Test, and Logistic Regression when the proportion of DIF items is large*. American Educational Research Association.

Gonzales, A., Padilla, J.L., Dolores, H., Gomez-Benito, J., & Benitez, I. (2010). EASY-DIF: Software for analyzing differential item functioning using the Mantel-Haenszel and Standardization procedures. *Applied Psychological Measurement*. doi:10.1177/0146621610381489.

Graham, J.W. (2009). Missing Data Analysis: Making it work in the real world. *Annual Review of Psychology*, *60*(4), 549-576.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, *70*(5), 646-675.

Gözen Çıtak, G. (2007). *Klasik test ve madde-tepki kuramlarına göre çoktan seçmeli testlerde farklı puanlama yöntemlerinin karşılaştırılması.* Doktora Tezi, Ankara Üniversitesi, Ankara

Hambletton, R.K. & Swaminathan, H. (1985). *Item Response Theory: Principles and applications*. Boston: Kluwer-Nijhoff Publishing.

Harwell, M. Stone, C. A., Hsu, T.C., & Kirisci, L. (1996). Monte carlo studies in item response theory. *Applied Psychological Measurement*, 20, 101-125.

Hohensinn, C. & Kubinger K. D. (2011). On the impact of missing values on item fit and the model validness of the Rasch model. *Psychological Test and Assessment Modeling*, 53, 380-393.

Kan, A., Sünbül, Ö., Ömür, S. (2013). 6.- 8. sınıf seviye belirleme sınavları alt testlerinin çeşitli yöntemlere göre değişen madde fonksiyonlarının incelenmesi. *Mersin University Journal of the Faculty of Education*, *9*(2), 207-222.

Kothari, C.R. (2004). *Research methodology: Methods and techniques (Second Revised Edition).* New Delhi: New Age Int. Ltd.

Kristanjansonn E., R. Aylesworth, I. McDowell & B.D. Zumbo (2005). A Comparison of four methods for detecting differential item functioning in ordered response model. *Educational and Psychological Measurement. 65*(6), 935-953.

Little, R. J. A & Rubin, D. B. (1987). *Statistical analysis with missing data (2nd ed.).* New York: John Wiley & Sons, Inc.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika*, 39, 247-264.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. New Jersey: Lawrence Erlbaum Associates.

Molenberghs, G., & Kenward, M.G. (2007). *Missing data in clinical studie (1 st ed.)*. England: John Wiley&Sons.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and Simultaneous Item Bias procedures for detecting differential ıtem functioning, *Applied Psychological Measurement, 18*(4).

Osterlind, S.J. (1983). *Test item bias*. London: Sage Publication.

Padilla, J.L., Hidalgo, J.L., Benitez, I., & Gomez-Benito, J. (2012). Comparison of three software programs for evaluating DIF by means of the Mantel-Haenszel procedure; EASY DIF, DIFAS and EZDIF, *Psicologica*, 33,135-156.

Peng, C.Y.J., Harwell, M., Liou, S.M., & Ehman, L. H. (2006). *Advances in missing data methods and implications for educational research. In S. Sawilowsky* (Ed.), Greenwich: Real data analysis.

Peng, C. J., & Zhu, J. (2008). Comparison of two approaches for handling missing covariates in logistic regression. *Educational and Psychological Measurement, 68(1),* 58-77.

Pigott, T.D. (2001). A review of methods for missing data. *Educational Research and Evaluation, 7*(4); 353-383.

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.

Rousseau, M., Bertrand, R., & Boiteau, N. (2006, April). *Impact of missing data treatment on the efficiency of DIF methods.* California: National Council on Measurement in Education.

Royce, S., Straits, B.C., & Straits, M.M. (1993). *Approaches to social research* (2nd ed.). New York: Oxford University Press.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581-592.

Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research,* (8), 3-15.

Sedivy, S. K., Zhang, B., & Traxel, N. M. (2006). *Detection of differential item functioning with polytomous items in the presence of missing data.* California: National Council of Measurement in Education.

Selvi, H. (2013). *Klasik test ve madde tepki kuramlarına dayalı değişen madde fonksiyonu belirleme tekniklerinin farklı puanlama durumlarında incelenmesi.* Yayınlanmamış Doktora Tezi. Mersin Üniversitesi Eğitim Bilimleri Enstitüsü.

Singh, Y.K. (2006). *Fundamental of research methodology and statistics*. New Delhi: New Age Int. Ltd.

Spray, J., & Miller, T. (1994). *Identifying nonuniform DIF in polytomously scored test items.* American College Testing Research Report Series 94-1. Iowa City, IA: American College Testing Program.

Ward, W.C., & Bennett, R.E. (2012). *Construction versus choice in cognitive measurement: issues in constructed response, performance testing, and portfolio assessment.* London and New York: Routledge, Taylor & Francis Group.

Woodward, M., Smith, W.C., & Tunstall Pedoe H. (1991). Bias from missing values: Sex differences in implication of failed venepuncture for the Scottish Health Study.*Int J. Epidemiol*.

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research & Evaluation*, *12*(3), 1-26.

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.