ISSN: 2790-4407

Journal of Computer Adaptive Testing in Africa

2022; 1, 29-39

Detection of Differential Item Functioning Magnitude in Psychological Measurement with Missing Data

[a]Alexander Olayinka Oluwafemi , [b]Femi Timothy Adekunle* & [c]Eyitayo Rufus Ifedayo Afolabi

[a,b,c] Department of Educational Foundations and Counselling, Obafemi Awolowo University, Nigeria

ABSTRACT

The paper investigated the effectiveness of missing data methods in detecting differential item functioning magnitude in polytomous scored non-cognitive items with a view to determining different levels of magnitude existing in non-cognitive items and the difference in the ability of missing data methods to detect DIF magnitude. Using the sample of 1,500 senior secondary school students, drawn through multistage sampling technique from Osun State, data were collected with the Achievement Motivation Inventory (AMI). The result showed that with Full Information Maximum Likelihood (FIML) 81.3% possess small DIF magnitude, while 4 (12.5%) items possess moderate DIF magnitude, while high DIF magnitude occurs in 2(6.3%) items, and 24 (75.0%) were categorised as having small DIF magnitude, 2 (6.3%) with moderate DIF magnitude, while 6 (18.8%) was classified as having high magnitude of DIF using Multiple Imputation (MI). The result further revealed that item effect size (magnitude) across different methods did not distinctly differ from one another ($X^2 = 0.16$, df = 1, $p > 0.05$). The study concluded that the two missing data methods were effective in detecting magnitudes of DIF present in polytomous scored non-cognitive items, and the differences that exist in the abilities of these missing data methods is not statistically significant.

Keywords: Achievement motivation inventory, differential item functioning, full information maximum likelihood, multiple imputation

* Corresponding author.

Department of Educational Foundations and Counselling, Obafemi Awolowo University, Nigeria, e-mail: femalex2003@yahoo.com

To cite this article:

Oluwafemi, A. O., Adekunle, F. T. & Afolabi, E. R. I. (2022). Detection of Differential Item Functioning Magnitude in Psychological Measurement with Missing Data. *Journal of Computerized Adaptive Testing in Africa, 1*, 29-39.

Received: May 23 2021
Accepted: November 29 2021
Published: June 27 2022

INTRODUCTION

In social science and educational research, differential item functioning (DIF) has been described as one of the factors that affect the validity of a test which may also result in bias estimates. The presence of DIF makes it difficult to make valued inference or deductions from the results of a test or study. Differential item functioning occurs when the expected item score conditioned on the latent trait differs due to group membership. Differential item functioning (DIF) analyses can provide evidence of the degree to which scores are comparable across groups. It is evident when a test item has different psychometric characteristics for members of different groups despite there being no difference in their overall ability on what is being measured. In DIF terminology the studied groups are referred to as focal and reference groups. The reference group is typically a majority group against whom the focal group is compared. The focal group may be believed to have potential educational or societal disadvantage whereas the reference group may be believed to have a relative advantage.

DIF magnitude otherwise known as DIF effect size describes the amount of differential item functioning present in a test item. It describes the significance of bias estimates that may be detected in the test item under study. Garret (2009) described DIF magnitude as effect size measures which are statistical tools used to determine the practical significance of DIF. Many times, items may be flagged as possessing statistically significant DIF, but the item contains a small amount of DIF. Small amounts of DIF may not have much impact on examinee or respondent scores, whereas large amounts of DIF will likely have more impact on examinee or respondent scores.

For a psychological test to function properly as intended, items in the test should measure respondents' performance fairly across different groups of respondents such as male and female. One of the core issues in comparing individuals and groups is to ensure that item bias is investigated in order to minimize inappropriate interpretations. When tests are labeled "biased", the accusations often have to do with the instruments chosen for a particular context, the way in which these tests are administered or the way in which the results are interpreted and/or used.

In educational testing, missing data occur when a respondent either does not respond to an item or question (i.e., item non-response) or does not respond to any question at all (i.e., unit non-response). That is, data are missing for some test items, and / or for some respondents. When students do not

answer items in a test because they do not know the answer, do not have time to respond to all questions, or omit the questions they are not comfortable with (such as in the case of attitudinal measurement), the item non-response generates a missing data problem (e.g., the variable of interest and the omitted response are not independent) which cannot be ignored (i.e., leaving data untreated, doing nothing about it). Orley (2017) describe missing data in the context of an examinee who may simply run out of time before reaching the item, or skip an item with the intention of returning to answer it later only to run out of time, or forget that he skipped it.

Rubin (1976) described three probabilistic explanations for why data are missing. These include missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Data are MCAR if there is no justifiable reason for why it is missing i.e. randomness explains the missing data. A test could have MCAR data if both focal and reference examinees accidently skipped some items. Data may be MAR if the chance of omitting it is systematically related to data that has been observed. For example, in a 30- item test where Item 1 is the studied item, examinees missing response to Item 1 could be attributed to their group membership (focal, reference), and/or their observed performance on Items 2-29. Data are MNAR if the likelihood of it being missing is systematically related to data that has not been observed. Using the previous illustration, examinees missing response to Item 1 could be attributed to their potential performance on Item 1.

Some common traditional missing data techniques include list-wise deletion(also called complete-case analysis or case-wise deletion), in which cases with missing values are discarded, so the analyses are restricted to cases that have complete data. The major advantage of listwise deletion is that it produces a complete dataset, which in turn allows for the use of standard analysis techniques. However, the disadvantages are numerous. Not surprisingly, deleting incomplete records can dramatically reduce the total sample size, particularly for data sets that include a large proportion of missing data or many variables. As a result, significance tests will lack power. More importantly, list-wise deletion assumes that the data are MCAR (i.e., missingness is unrelated to all measured variables). When the MCAR assumption is violated – as it often is in real research settings – the analyses will produce biased estimates.

Pairwise deletion approach (also known as available-case analysis) is another commonly used deletion technique which is often regarded as an improvement over listwise deletion because it minimizes the number of cases discarded in any given analysis, but it still suffers from the same major limitation as listwise deletion, namely that the data are MCAR. Like listwise deletion, pairwise deletion can produce

biased estimates when the data are inconsistent with an MCAR mechanism, Mean imputation replaces missing values with the arithmetic mean of the available data. It maintains the sample size and it is easy to use but the variability in the data is reduced causing the standard deviation and the variation to be under estimated. Regression imputation retains individuals for all statistical analyses by predicting their missing data values from a linear regression equation that is constructed from observed variables in the dataset. In a bivariate analysis with missing data on a single variable, the complete cases are used to estimate a regression equation where the incomplete variable serves as the outcome and the complete variable is the predictor. The resulting regression equation generates predicted scores for the incomplete cases.

The common modern missing data methods presently in use include Maximum likelihood estimation and multiple imputation which are considered state of the art missing data techniques and are widely recommended in the methodological literature (Baraldi & Enders, 2009; Schafer & Graham, 2002; Rodriguez De Gil, 2015). These approaches are superior to traditional missing data techniques because they produce unbiased estimates with both MCAR and MAR data. In this study, two widely used missing data methods i.e. Full Information Maximum Likelihood (FIML) and Multiple Imputation (MI) were used.

From the fore-going submissions, two research questions and one hypothesis were postulated:

Research Questions

(i) What is the pattern of differential item functioning magnitude detected with Full Information Maximum Likelihood (FIML)?

(ii) What is the pattern of differential item functioning magnitude detected with Multiple Imputation (MI)?

Hypothesis

(i) There is no significant difference in the magnitude of differential item functioning detected across missing data methods

MATERIAL AND METHODS

The study adopted a descriptive survey design. The population consisted of all senior secondary school students in Osun state. A sample of 1500 senior secondary school III students were selected in the state across the three senatorial districts using multistage sampling techniques. From each of the

educational districts five Local Government Areas (LGAs) were selected using simple random sampling technique. From each of the selected LGAs, two schools which had not less than 50 senior school III students were purposively selected. From each of the sampled schools, the students in Senior Secondary School III were purposively selected. This is because they are expected to have better achievement motivation than their other colleagues as a result of their presumed level of preparedness for their final examination in the school. Every Senior Secondary School III students in the selected schools were used so as to avoid the problem of keeping some students out of the classroom

Research Instrument

The instrument used for data was an adapted version of Achievement Motivation Inventory (AMI) developed by Muthee and Thomas (2009). The instrument which measured a broad construct of academic related achievement motivation has 32 Likert type items and was pretested for validity and reliability checks to ensure that it is suitable for data collection using 100 senior secondary school III students who were not part of the study sample. The reliability was established using Kuder-Richardson method which produces 0.780 index of internal consistency.

Methods of Data collection

The process of data collection for the study started with a visit by the researchers to the Local Inspectors of Education in the selected local government areas to ascertain the number of students in the Senior Secondary School III in the LGAs. The researchers then moved to the schools to seek permission for the use of their students in the study. This was facilitated after an administrative letter requesting for permission to carry out the study had been submitted to the administrative head of the schools being used for the study. The Achievement Motivation Inventory was administered by the researcher with the assistance of class teachers and research assistants while the students were reminded that their participation is voluntary and they were not to identify themselves on the questionnaire

RESULTS

Research Question One: What is the pattern of differential item functioning magnitude detected with Full Information Maximum Likelihood (FIML)?

In answering this question, analysis was conducted on measures based on the group differences in the expected item scores which connote the chi-square value for each of the item. This is otherwise known

as item level effect sizes of DIF (i.e. magnitude). However, Thissen, Steinberg, & Wainer (1993) and Awour (2008) classified the differences of effect size as small when chi-square value (G^2) ≤ 1, moderate when 1 <G^2≤ 1.5 and high G^2 ≥ 1.5 respectively.

Table 1: Magnitude of DIF statistics with respect to FIML

	AIC	AICc	SABIC	HQ	BIC	G^2	Magnitude
item1	2.00	2.91	4.13	3.98	7.31	0.00	Small
item2	1.96	2.87	4.10	3.94	7.27	0.04	Small
item3	1.94	2.85	4.07	3.92	7.25	0.06	Small
item4	1.13	2.04	3.26	3.10	6.44	0.88	Small
item5	1.82	2.73	3.95	3.80	7.13	0.18	Small
item6	1.98	2.89	4.11	3.96	7.29	0.02	Small
item7	0.77	1.68	2.90	2.75	6.08	1.23	Moderate
item8	1.35	2.26	3.48	3.33	6.66	0.65	Small
item9	1.80	2.71	3.93	3.77	7.11	0.21	Small
item10	1.78	2.69	3.91	3.76	7.09	0.22	Small
item11	-0.90	0.01	1.24	1.08	4.41	2.90	High
item12	1.93	2.84	4.06	3.91	7.24	0.07	Small
item13	0.79	1.70	2.92	2.77	6.10	1.21	Moderate
item14	1.97	2.88	4.10	3.95	7.28	0.03	Small
item15	1.82	2.73	3.95	3.79	7.13	0.18	Small
item16	1.94	2.85	4.07	3.91	7.25	0.07	Small
item17	1.72	2.63	3.85	3.70	7.03	0.28	Small
item18	1.69	2.60	3.83	3.67	7.00	0.31	Small
item19	2.00	2.91	4.13	3.98	7.31	0.00	Small
item20	1.79	2.70	3.93	3.77	7.10	0.21	Small
item21	0.79	1.70	2.93	2.77	6.10	1.21	Moderate
item22	0.86	1.77	2.99	2.84	6.17	1.14	Moderate
item23	1.82	2.74	3.96	3.80	7.13	0.18	Small
item24	-0.86	0.05	1.27	1.12	4.45	2.86	High
item25	1.28	2.19	3.41	3.26	6.59	0.72	Small
item26	1.99	2.90	4.13	3.97	7.30	0.01	Small
item27	1.77	2.68	3.90	3.75	7.08	0.23	Small
item28	1.89	2.80	4.03	3.87	7.20	0.11	Small
item29	1.74	2.65	3.87	3.72	7.05	0.26	Small
item30	1.90	2.81	4.03	3.88	7.21	0.10	Small
item31	1.96	2.87	4.10	3.94	7.27	0.04	Small
item32	1.79	2.71	3.93	3.77	7.11	0.21	Small

Table 1 revealed that majority of the items (26 items) amounting to 81.3% possesses small DIF magnitude while 4 (12.5%) items possesses moderate DIF magnitude while high DIF magnitude occurs in 2(6.3%) items. This is implies that Full Information Maximum Likelihood (FIML) was able to detect different categories of DIF magnitude across different subgroups.

Research Question Two: What is the pattern of differential item functioning magnitude detected with Multiple Imputation (MI)?

In answering this question, the classification of DIF magnitude as advanced by Educational Testing Service and reported by Thissen et al. (1993); Awour (2008) was used.

Table 2: Magnitude of DIF statistics with respect to MI

	AIC	AICc	SABIC	HQ	BIC	G^2	Magnitude
item1	1.96	2.87	4.09	3.94	7.27	0.04	Small
item2	1.98	2.89	4.11	3.95	7.29	0.03	Small
item3	1.96	2.87	4.09	3.94	7.27	0.04	Small
item4	0.69	1.60	2.82	2.67	6.00	1.31	Moderate
item5	1.74	2.65	3.87	3.72	7.05	0.26	Small
item6	1.99	2.90	4.12	3.97	7.30	0.01	Small
item7	1.38	2.29	3.52	3.36	6.69	0.62	Small
item8	1.70	2.61	3.83	3.68	7.01	0.30	Small
item9	1.94	2.85	4.07	3.91	7.25	0.07	Small
item10	1.87	2.78	4.00	3.85	7.18	0.13	Small
item11	-0.27	0.64	1.87	1.71	5.04	2.27	High
item12	1.82	2.73	3.96	3.80	7.13	0.18	Small
item13	0.14	1.05	2.27	2.12	5.45	1.86	High
item14	1.85	2.76	3.98	3.83	7.16	0.15	Small
item15	1.90	2.81	4.03	3.87	7.21	0.10	Small
item16	1.99	2.90	4.12	3.96	7.30	0.01	Small
item17	1.99	2.90	4.12	3.97	7.30	0.01	Small
item18	1.61	2.52	3.74	3.59	6.92	0.39	Small
item19	2.00	2.91	4.13	3.98	7.31	0.00	Small
item20	2.00	2.91	4.13	3.98	7.31	0.00	Small
item21	-0.01	0.91	2.13	1.97	5.30	2.01	High
item22	-0.90	0.01	1.23	1.08	4.41	2.90	High
item23	1.96	2.87	4.09	3.94	7.27	0.04	Small
item24	-1.23	-0.32	0.91	0.75	4.08	3.23	High
item25	-0.45	0.46	1.69	1.53	4.86	2.45	High
item26	1.88	2.79	4.01	3.86	7.19	0.12	Small
item27	0.88	1.80	3.02	2.86	6.20	1.12	Moderate
item28	1.95	2.86	4.08	3.92	7.26	0.06	Small

item29	1.96	2.87	4.09	3.93	7.27	0.05	Small
item30	1.90	2.81	4.04	3.88	7.21	0.10	Small
item31	1.67	2.59	3.81	3.65	6.98	0.33	Small
item32	1.89	2.80	4.02	3.87	7.20	0.11	Small

From Table 2, it was observed that there are 24 (75.0%) categorized as having small DIF magnitude, 2 (6.3%) with moderate DIF magnitude while 6 (18.8%) was classified as having high magnitude of DIF. This is implies that Multiple Imputation (MI) was able to detect different categories of DIF magnitude (DIF item effect size) across different subgroups.

Tables 1 and 2 revealed differential item functioning level of magnitude (That is it describes the amount of DIF present in a particular item) for each of item across all the methods with respect to sub group. It was observed that for FIML, the differences had 26 (81.3%) as small item effect size, 4 (12.5%) as moderate while 2(6.3%) was classified as high magnitude of DIF. Moreover, a critical examination of Tables 2 depicts that MI differences had 24 (75.0%) as small item effect size, 2 (6.3%) as moderate while 6 (18.8%) was classified as high magnitude of DIF. It was inferred from this results that, in overall, many of the items across the two methods were of small item effect size (that is magnitude) and items with moderate and high effect size (magnitude) were very few respectively.

Hypothesis

There is no significant difference in the magnitude of differential item functioning detected across missing data methods

In answering this question, normality assessment of dataset using normal P-P plot revealed a non-normal distribution. This suggested that non-parametric statistical tool should be used. Consequently, Kruskal-Wallis one-way analysis of variance (ANOVA) of non-parametric method for comparing k independent samples was used to test for the significance difference in the magnitude of DIF across missing data methods. This is roughly equivalent to a parametric one-way ANOVA with the data replaced by their ranks. The table below presents Kruskal-Wallis statistics across missing data methods with respect to sub-group respectively.

Table 3: Kruskal-Wallis Statistics for Magnitude across Methods with respect to sex

Null Hypothesis	Chi-Square	df	Asymp. Sig.	Decision
Magnitude across methods with respect to sex are the same	0.16	1.00	0.98	Null hypothesis was not rejected

Table 3 shows the significance of differences observed in the magnitude across missing data methods with respect to sex. Hypothesis was tested using Kruskal-Wallis of One-Way Analysis of Variance by Ranks. The result shows that the stated null hypothesis that there is no significance difference in the magnitude across methods with respect to sex was not rejected with (Chi-square (X^2) = 0.16, df = 1, $p > 0.05$). This implies that item effect size (magnitude) across different methods were not distinctly differs from one another.

DISCUSSION

Magnitude measures are an essential part of DIF detection because of the need to avoid false positives particularly in an environment in which items have been studied carefully and subjected to qualitative and quantitative analyses prior to DIF detection. It is desirable to identify and flag only items with salient DIF. From the result of this study, it was observed that while majority of the items displayed small DIF magnitude, very few of the items shows moderate and high DIF magnitude across missing data methods. According to Zwick (2012) magnitude of DIF can be placed in three categories with categories A (*negligible or non-significant DIF*), B (*slight to moderate DIF*), or C (*moderate to large DIF*) and that this DIF classification which was designed by education testing service (ETS) over the years have undergone certain modifications but the overall classification still remain intact. This categories was further broken down by Monahan, Mchorney, Stump and Perkins (2007) who reported that

Category A. Items with negligible or non-significant DIF. Defined as not significantly different from zero or absolute value less than 1.0.

Category B. Items with slight to moderate magnitude of statistically significant DIF. Defined as different from zero and absolute value of at least 1.0 and either less than 1.5 or not significantly greater than 1.0.

Category C. Items with moderate to large magnitude of statistically significant DIF. Defined by absolute value of at least 1.5 and significantly greater than 1.0.

The study further tested the significant difference in the ability of missing data methods under review to detect and categorize DIF magnitude. It found no significant difference in the ability of these methods to detect and categorize DIF magnitude in polytomous items. Robitzsch and Rupp (2009), reported that the amount of bias that got introduced in some cases was of the order of magnitude of large DIF effect sizes when an incorrect method of missing data (i.e., an approach where incorrect

values are imputed) is employed and this can either induce DIF when no DIF is present (i.e., result in biases and inflated type-I error rates) or mask DIF when it is, indeed, present (i.e., result in biases and reduced power rates). The result is consistence with the findings of Kilmen (2016) and Garret (2009) who reported changes in DIF magnitude as the power to detect DIF irrespective of whether missing data are present or not.

CONCLUSION

The study concluded that it is not enough for DIF to be detected in an item but further analysis should be made to indicate the magnitude of DIF present in such an item to ascertain whether the items possess DIF of statistical significance. The two missing data were effective in detecting Magnitudes of DIF present in polytomous scored non cognitive item and the differences which exists in the abilities of these missing data methods is not statistically significant.

REFERENCES

Awuor, R. A. (2008). Effect of Unequal Sample Sizes on the Power of DIF Detection: An IRT Based Monte Carlo Study with SIBTEST and Mantel-Haenszel Procedures. *Unpublished Ph.D Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University.*

Baraldi N. A., & Enders, C.K. (2009).An introduction to modern missing data analyses. *Journal of School Psychology* 48 (2010) 5–37

Garrett, P. L. (2009). "A Monte Carlo Study Investigating Missing Data, Differential Item Functioning, and Effect Size." *Dissertation, Georgia State University, 2009.* https://scholarworks.gsu.edu/eps_diss/35

Kilmen, S. (2016). Effect of DIF Magnitudes, Focal Group Sample Size, and DIF Ratio on the Performance of SIBTEST. *International J. Soc. Sci. & Education* 2016 Vol.6 Issue 1,ISSN: 2223- 4934 E and 2227-393X Print

Monahan, Mchorney, Stump and Perkins (2007). Odds Ratio, Delta, ETS Classification, and Standardization Measures of DIF Magnitude for Binary Logistic Regression. *Journal of Educational and Behavioral Statistics* Vol. 32, No. 1, pp. 92–109 DOI: 10.3102/107699860629803-AERA and ASA. http://jebs.aera.net

Muthee, J. M. & Thomas, I. (2009). Predictors of Achievement Motivation Among Kenyan Adolescents *the Psyche space* Vol. 3, No. 2, 39-44, July 2009.

Orley, G. J., Multiple Imputation of the Guessing Parameter in the Case of Missing Data (2017). *Public Access Theses and Dissertations from the College of Education and Human Sciences.* 288. Retrieved from http://digitalcommons.unl.edu/cehsdiss/288

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.

Rodriguez De Gil, P. (2015). An Empirical Comparison of the Effect of Missing Data on Type I Error and Statistical Power of the Likelihood Ratio Test for Differential Item Functioning: An Item Response Theory Approach using the Graded Response Model. *An unpublished Ph.D Thesis of the University of South Florida*

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art.*Psychological Methods, 7*(2), 147-177

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models.In P. W. Holland, & H. Wainer (Eds.) *Differential item functioning,* (pp. 130-215). Hillsdale, England: Lawrence Erlbaum Associates, Inc.

Zwick, R. (2012). A Review of ETS Differential Item Functioning Assessment Procedures: Flagging Rules, Minimum Sample Size Requirements, and Criterion Refinement. http://www.ets.org/research/contact.html