American Journal of Applied Psychology
2016; 5(6): 38-43
http://www.sciencepublishinggroup.com/j/ajap
doi: 10.11648/j.ajap.20160506.11
ISSN: 2328-5664 (Print); ISSN: 2328-5672 (Online)

Comparison of DIF Detection Performances of Mantel Test and Likelihood Ratio Test

Safiye Bilican Demir

Department of Educational Sciences, Kocaeli University, Kocaeli, Turkey

Email address:
safiyebilican@gmail.com

To cite this article:
Safiye Bilican Demir. Comparison of DIF Detection Performances of Mantel Test and Likelihood Ratio Test. *American Journal of Applied Psychology.* Vol. 5, No. 6, 2016, pp. 38-43. doi: 10.11648/j.ajap.20160506.11

Received: October 20, 2016; **Accepted**: November 3, 2016; **Published**: November 25, 2016

Abstract: The purpose of this study was to investigate Type I error rate of the IRT-Likelihood Ratio (IRT-LR) statistic and Mantel Test in detecting DIF. A multiple replication Monte Carlo study was utilized for simulated data sets. In final study design, there were 18 conditions [3 (sample size) x 3 (group mean difference) x 2 (methods of DIF detection)]. WinGen3 was used to simulate ability estimates and to generate response data sets. MULTİLOG and DIFAS were used to conduct the Mantel and IRT-LR DIF analyses. Results indicated that with equal group distribution, Mantel Test and IRT-LR Test performed similarly under all testing conditions and had better Type I error rate control. Large sample size and presence of group mean difference tended to inflate the Type I error rates of both DIF detection tests. IRT-LR had higher Type I error rates than Mantel Test when large sample size and when group mean difference conditions.

Keywords: Differential Item Functioning, Monte Carlo, Polytomous Items, Type I Error

1. Introduction

Tests are used in many areas to make decisions about individuals. Based on the test results, important decisions on different subjects like placing of individuals in an educational institution, determining individuals' academic or work performances, personnel selection, tracking learning, giving feedback, diagnosing behavioral disorders and professional guidance are taken. The decisions taken may have significant effects on the individuals' personal, social and political situations. In this case, since the qualifications of measurement instruments affect these decisions, scores obtained from a measurement instrument must be reliable, comparable and fair [1].

To determine a measure's validity, assessment developers and researchers often conduct validation studies. These studies are conducted to assemble evidence regarding the strength of an assessment's inferences [2]. If the assembled evidence indicates a measure has strong inferences, then that measure is labeled highly valid. A measure's validity, however, can be adversely affected by a number of factors. One such factor is differential item functioning (DIF). DIF occurs when an item performs differently for two contrasting groups of respondents (e.g., male vs female) after controlling for differences in the abilities of the groups [3]. The unexpected performance difference may cause inaccurate trait inferences for certain examinees, in turn, adversely affecting test validity.

Ensuring that tests do not contain DIF has become an important part of developing valid assessments. Methods for detection of DIF have grown, in large part, due to the legal and ethical need to measure respondent performance without bias [4]. While DIF detection is predominantly used in the cognitive context, where answer choices are usually dichotomous, it can also be used in areas where items are typically polytomously scored [5]. There has been a marked increase in the use of polytomous assessments in education. The use of open-ended, performance assessments, performance task and constructed-response instruments to assess educational outcomes have greatly increased during the last decade [6]. In recent years, nationwide testing and assessment programs have included polytomously scored items in their assessments [7].

Two settings in which polytomous items are frequently used are ability assessment and attitude assessment. Ability assessments measure cognitive traits such as reading comprehension, written expression, or math. These assessments can be comprised entirely of polytomous items

or contain a mixture of dichotomous and polytomous items [8]. Unlike ability assessments, attitude assessments usually contain only polytomous items. Their widespread use in attitude assessment is the result of the type of traits these assessments measure. Polytomous items are vital in predilection measures because dichotomous responses of correct or incorrect are inappropriate responses to indicate feelings and opinions, and while dichotomous responses of true or false could be used for attitude assessment, they do not provide the appropriate range of choices for respondents. The increased information on the underlying trait that multiple response categories provide is one of the main reasons for the proliferation of polytomous item formats [9]. As a result, polytomous items are typically chosen for these types of assessments. Consequently, attitude assessments and many of these performance measures consist entirely of polytomous items rather than multiple choice items. Indeed, the use of polytomous item formats nationwide has led to increased attention to the detection of DIF in these items [10].

A variety of procedures for detecting possible item bias through DIF have been developed for polytomous items. Some of these methods are based on classical test theory (CTT). Mantel-Haenszel Test (MH) is largely used and can be given as examples ofthe methods based on CTT. Some DIF determination methods are based on item response theory (IRT) and likelihood ratio (IRT-LR) is example of this method. In this study, the performance of two popular DIF detection test Mantel Test and IRT-LR test was compared.

In the IRT-LR test for DIF detection [11], the null hypothesis to be tested is that the item parameters between the reference group and the focal group do not differ. For the test of the null hypothesis of no DIF, two models are compared: a compact model and an augmented model. In the compact model, the item parameters for the common item or items across groups are constrained to be equal in the two groups. In the augmented model, the item parameters for the studied item are unconstrained and the remaining items are constrained to be equal in the two groups. Then the likelihood-ratio test statistic, G^2. The value of G^2 is distributed as the chi-square with the degrees of freedom equal to the difference in the number of parameters in the two models. If the result of the test is found to be significant, then it is said that the studied item exhibit DIF.

The Mantel Test, a nonparametric observed score method, is a polytomous extension of the Mantel-Haenszel method that takes into account the ordered nature of the response categories when testing for DIF. The Mantel Test provides a statistic with a chi-square distribution of one degree of freedom when the null hypothesis of no DIF is true [12]. Calculation is based on item means for groups that have been matched on some measure of proficiency. Because Mantel Test is conceptually simple, does not require large sample size, and provides a chi-square test of significance, it has become a widely used method for detecting DIF.

Simulation studies investigating the performance of DIF detection methods for polytomous items have increased over the years. The use of simulated data in the current research is highly desirable for two reasons. First, with Monte Carlo methods, researchers knows the true item parameters that were used to simulate the data and thereby controls which items contain DIF. This allows one to compare the results of the DIF anaysis to the true characteristics of the data. Second, Monte Carlo methods allow the researcher to manipulate other characteristics of the data as well, such as sample size. One can examine the factors moderating a statistics ability to detect DIF. Because this study aims to assess the efficacy of DIF detection methots in different conditions, a Monte Carlo research design is necessary as it provides only means possible to identify true DIF items and measures the accuracy of DIF detection.

To reduce the potential threats of DIF to test validity, it is important to determine the effectiveness of DIF detection methods for polytomous items in the different sample size and gropu ability distrubition. For the IRT-LR and Mantel Test that are usually preferred in actual test applications, under which test conditions these tests show better performance or under which conditions these tests are more sensitive must be determined by comparing performances in the context of Type I error error under changing test conditions. In this case, knowing how much these IRT detection techniques give more consistent results under changing conditions will make test results more valid for test administrators and developers.

The purpose of this study was to investigate the Type I error rate of the likelihood ratio statistic and Mantel Test in detecting DIF in different sample size and group ability distribition conditions.

2. Method

2.1. Research Design

A multiple-replication simulation study was employed to evaluate the performance of the LR statistic and the Mantel approach when the properties of the data are known.

2.2. Data Simulated

The polytomous scored data were generated for the reference and the focal groups with a one-parameter IRT model. The Type I error conditions had factors that were held constant and factors that were varied:

2.2.1. Factors Held Constant

i. Polytomous IRT model. The Partial Credit Model (PCM) was used to generate the data for the reference group and the focal group. This model has been used in many simulation studies on DIF studies [e.g 10, 13].

ii. Test length There were 20 items generated under the PCM. This is a common test length in simulation studies investigating DIF. This test length is also similar to DIF detection studies investigating the impact of DIF for attitude and cognitive items [e.g 13, 14]

iii. Number of item categories. Each item was generated to

have four score categories (i.e., one point for each correct step) to simulate four ordered levels of performance that an examinee must execute in order to arrive at the correct solution to the problem.

iv. Type of DIF. Uniform DIF was the only type of DIF investigated in this study since the PCM has only δ *b*-parameters.

2.2.2. Factors Varied

i. sample size. Three levels of sample size were investigated, with the number of examinees in the reference/focal groups being 250/250, 1000/250, and 1000/1000. The sample size of 2,000 was chosen to ensure that item parameter estimation errors would be minimized in the MULTILOG calibrations, and the sample size of 500 was chosen so that the effect of small sample size on the performance of the IRT-LR statistic could be studied. The 1000/250 sample size condition was chosen to represent the case in which the sample size of the focal group is substantially smaller than that of the reference group [15], a situation commonly encountered in testing situations when the reference and focal groups are defined by ethnicity.

ii. Ability distribution differences. There were three levels of between-group differences in ability distribution investigated in this study. When the reference and focal groups had the same ability, the data were generated to have a mean of 0 and a standard deviation of 1 (R~*N* (0, 1), O~*N* (0, 1)). When ability distributions between the focal group and the reference group differ, the data for the focal group were generated to have a mean of -0.5 and -1 and a standard deviation of 1 ((R~*N* (0, 1), O~*N* (-0.5, 1)) - (R~*N* (0, 1), O~*N* (-1, 1)), whereas the data for the reference group were generated to have a mean of 0 and a standard deviation of 1.

In finally study design, The Type I error portion of this study involved conditions where no DIF was present. Three factors were fully crossed: 3 (group ability differences) x 2 (sample size) x 2 (DIF detection tests) = 18 fully crossed conditions.

In this study, a total of 100 replications were completed for each condition. This number of replications is consistent with much previous researh [e.g., 16, 17, 18] Responses for the reference group and focal group members were generated separately and were later combined to create one data set containing both reference group and focal group responses.

2.3. Data Analysis

For the IRT-LR test of the null hypothesis of no DIF, two models are compared: a compact model and an augmented model. In the compact model, the item parameters for the common item or items across groups are constrained to be equal in the two groups. In the augmented model, the item parameters for the studied item are unconstrained and the remaining items are constrained to be equal in the two groups. Then the likelihood-ratio test statistic, G^2, is computed. The value of G^2 is distributed as the chi-square with the degrees of freedom equal to the difference in the number of parameters in the two models. If the result of the test is found to be significant, then it is said that the studied item exhibit DIF.

In Mantel test, mantel chi-square statistics was used to decide whether the relevant items show DIF or not. Total scores were used to match reference and focal group individuals. Chi-square statistics obtained after the data analysis showed chi-square distribution in a degree of freedom. For this statistics, critical value was 3.84 at the 0.05 significance level.

After the DIF analyses were completed, the performance of each DIF detection method was compared across all conditions to determine the Type I error of each DIF method under the various study conditions. Type I error was the proportion of times out of 100 replications where DIF was falsely identified at the 0.05 level.

In present study, Bradley's liberal criterion was used for a criterion of a Type I error. If a probability of Type I error falls within the criterion of 0.025 ≤ Probability of Type I error≤ 0.075 at nominal α level of 0.05 and 0.0055 ≤ Type I error ≤ .015 at nominal α level of. 01 [19]. The computer program MULTILOG was used for the IRT- LR test. The results for the MH were obtained using DIFAS.

3. Results and Discussion

3.1. Type I Error Rates for Mantel Test and IRT-LR for Different Sample Size in R~N (0,1), O~N (0,1)

The results for the Type I error rates for Mantel Test and IRT-LR are displayed in Table 1 for different sample size conditions including reference and focal group have same ability distribution.

Table 1. *Type I error rates for Mantel Test and IRT-LR in R~N (0,1), F~N (0,1) condition.*

Ability distribution		Sample size		DIF detection test	
Reference	Focal	Reference	Focal	Mantel Test	IRT-LR
R~N (0,1), O~N (0,1)		250	250	0.045	0.052
		1000	250	0.047	0.038
		1000	1000	0.053	0.041

According to Table 1, In R~N (0,1), O~N (0,1) condition, the Type I error rates for the Mantel Test Mantel and IRT-LR were at or close to the nominal rate of 5%. Type I error rates for the Mantel ranged from 0.045 to 0.053 5, whereas the Type I error rates for IRT-LR ranged from 0.038 to 0.052. The Type I error rates for both DIF detection methods were all close to the nominal rate of 0.05 in all conditions in which there were no mean latent trait differences between the reference and focal groups. Type I error rates not exceeded the nominal rate of 0.05 and fell within Bradley's (1978) liberal robustness criterion of 0.025 to 0.075 range for this study. It means that each DIF detection procedure provided adequate control of Type I error.

In related literature, there are many studies that compare Mantel Test and IRT-LR under different test conditions using Partial Credit Model or other multi-category scored IRT models (Graded response Model, Generalize Partial Credit

Model). These studies, also, showed that when the reference and focal group ability distributions exhibited normal distribution, Type I error ratios related to Mantel Test and IRT-LR were close to 0.05 α level or below it and controlled the Type I error [e.g 13, 16, 20, 21].

When no ability distribution differences, the Type I error rates increased for Mantel Test as the sample size increased while the Type I error rates for IRT-LR decreased. Mantel Test generally in small smaple size (250: 250) and IRT-LR in large sample size (1000:1000) condition maintained a better Type I error rate. This case may be explained as a way to obtain Mantel Test and IRT-LR's DIF statistics. It is known that Mantel Test that matches groups by observed scores evaluates item average difference between the groups based on the Chi-square statistics. Since this statistics is sensitive to increasing sample size, it increases the possibility of items with no DIF or very small amounts of DIF with increasing sample size identified as DIF. In this case, dependent on the increasing sample size Type I error ratios tend to increase for Mantel Test. IRT-LR matches the groups over a latent variable and evaluates the difference in of item parameters through model matching. The MULTILOG program used for the IRT-LR analysis uses marginal likelihood estimation method for the estimation of item parameters and maximum likelihood estimation method for the ability parameters. The number of individuals and ability distribution may play important roles in the accuracy of item and ability parameter estimations obtained from the groups [22]. In this regard, more accurate item and ability estimation depending on the increasing sample size may have contributed to IRT-LR giving lower Type I error ratios.

3.2. Type I Error Rates for Mantel Test and IRT-LR for Different Sample Size in Different Group Ability Distribution

Two levels of ability group distribution were investigated, with the reference/focal groups being F~N (-0.5, 1) and F~N (-1, 1).

3.2.1. Type I Error Rates for Mantel Test and IRT-LR for Different Sample Size in R~N (0, 1), F~N (-0.5, 1)

The results for the Type I error rates for Mantel Test and IRT-LR are displayed in Table 2 for different sample size conditions including reference and focal group have different ability distribution.

Table 2. *Type I error rates for Mantel Test and IRT-LR in R~N (0, 1), F~N (-0.5, 1) condition.*

Ability distribution		Sample size		DIF detection test	
Reference	Focal	Reference	Focal	Mantel Test	IRT-LR
R~N (0, 1), O~N (-0.5, 1)		250	250	0.043	0.069
		1000	250	0.048	0.071
		1000	1000	0.045	0.091

According to Table 2, In R~N (0, 1), F~N (-0.5, 1) condition, the Type I error rates for the Mantel Test ranged from 0.054 to 0.063, whereas the Type I error rates for IRT-LR ranged from 0.069 to 0.091. The Type I error rates for Mantel Test were all close to the nominal rate of 0.05 in all conditions. Type I error rates not exceeded the nominal rate of 0.05 and fell within Bradley's (1978) liberal robustness criterion of 0.025 to 0.075 range for this study. It means that Mantel Test provided adequate control of Type I error.

The Type I error rates for IRT-LR in all condition except large sample size condition fell within Bradley's (1978) liberal robustness criterion of. 025 to. 075. It showed that LR maintained a better Type I error rate in small and moderate sample size conditions. Also, for condition in which sample size is 1000:1000, IRT-LR began to lose over their average Type I error.

In all condition, Mantel Test had lower Type I error rates than IRT-LR. When ability distribution differences, the Type I error rates decreased for Mantel Test as the sample size increased while the Type I error rates for IRT-LR increased.

Compared to R~N (0, 1), O~N (0, 1) distribution conditions, while Type I error ratios decreased a far amount for Mantel Test, they increased for IRT-LR. Mantel Test produced lower Type I error ratios compared to IRT-LR depending on focal group's deviation in ability distribution average.

3.2.2. Type I Error Rates for Mantel Test and IRT-LR for Different Sample Size inR~N (0, 1), F~N (-1, 1)

The results for the Type I error rates for Mantel Test and IRT-LR are displayed in Table 3 for different sample size conditions including reference and focal group have different ability distribution.

Table 3. *Type I error rates for Mantel Test and IRT-LR in R~N (0, 1), F~N (-1, 1) condition.*

Ability distribution		Sample size		DIF detection test	
Reference	Focal	Reference	Focal	Mantel Test	IRT-LR
R~N (0, 1), O~N (-1, 1)		250	250	0.053	0.069
		1000	250	0.058	0.083
		1000	1000	0.063	0.122

According to Table 3, in R~N (0, 1), O~N (-1, 1) distribution condition, Type I error ratios ranged from 0.053 to 0.063 for Mantel Test and from 0.069 to 0.122 for IRT-LR. When Bradley's liberal criterion was taken into consideration, Mantel test controlled Type I error in all the sample size conditions and IRT-LR controlled Type I error in everything except the small sample size condition. In this condition, Type I error ratios increased depending on the increasing sample size for both of the DIF detection tests and the highest Type I error ratios were obtained in the big sample (1000:1000) condition where the deviation amount was -1. In all of the sample size conditions, Mantel Test produced lower Type I error ratio than the IRT-LR. The increase in groups' deviation amounts in ability average increased Type I error of IRT-LR the most.

The presence of group mean difference affected the Type I error results of both DIF detection tests. All in all, Type I error ratios for both of the DIF identification tests increased

depending on the deviation in the focal group ability distribution averages. The increase in Type I error ratios for Mantel Test may result from the differentiation of ability θ values expected in some of the raw score intervals depending on the change in focal groups' ability averages. Thus, observed item score average difference will be more apparent for certain score intervals.

Against the deviations in groups' ability distributions, IRT-LR remained weak in controlling Type I error and the likelihood of IRT-LR identifying an item that did not display DIF as an item that displayed DIF increased. This can be explained by the effect of deviations in θ distributions on the estimation of item parameters and by the fit between the item parameters and θ parameters.

In cases where the focal group ability distribution average is lower than the reference group, item parameters (step difficulty parameter) will take higher values correlatively with the ability distribution that is pulled down. In other words, step difficulty parameter dependent to the item can take lower values; it can take higher values for θ values. Accordingly, the error values will increase for parameter estimations obtained from the compact model where item parameters are matched in both of the groups and the model-data fit statistics obtained from this model will be worse compared to the augmented model because parameter estimations of the related item in the augmented model are made for each group separately so more accurate estimations will be obtained. Therefore, the augmented model that fits more may cause the related item to be marked as an item that displays DIF.

When we consider about fit between the item parameters and θ parameters, it is known that when the item parameter distribution gets closer to the θ parameters distribution, more accurate estimations are made [22]. The step difficulty parameters obtained in the study are obtained from the normal distribution where the average is 0 and standard deviation is 1. The deviations in focal group's θ distribution decrease the fit with the step difficulty parameters. Accordingly, the errors in parameter estimations depending on likelihood method and marginal probability method increase the most. This case can be explained as the cause of the observed difference among groups. In parallel to the findings of this study, in other studies where artificial data are used Type I error ratios for IRT-LR increased depending on the deviation in groups' ability distribution average and IRT-LR had difficulty in controlling the Type I error [13]

4. Conclusion and Recommendations

The study results showed that when the reference and focal group ability distributions exhibited normal distribution, both of the DIF detection tests controlled Type I error good. While the Type I error ratios of Mantel test increased for the condition that had similar group ability distributions depending on the increasing sample size, the ratios decreased for IRT-LR.

When deviations in focal group ability distribution averages come into question, Type I error ratios increased for both of the DIF identification tests depending on the increasing sample size and deviations in the focal group ability distribution averages. For both of the DIF identification tests, Type I error ratios increased depending on the increasing sample size and deviations in the focal group ability distribution averages. Compared to Mantel Test, the increase in Type I error values were higher for IRT-LR depending on the increasing sample size and deviations in the focal group ability distribution averages. For both tests, the highest Type I error was obtained from big sample size condition and increasing deviation amount in focal group's ability average condition. Compared to Mantel Test, the increase in Type I error values were higher for IRT-LR depending on increasing sample size and deviation in groups' ability average. Against the deviations in groups' ability distributions, IRT-LR remained weak in controlling Type I error.

According to the study results, when groups' ability distributions show normal distribution characteristic in analysis with polytomous items, it can be recommended for researchers to take the number of individuals in the reference and focal groups into consideration. They can use IRT-LR with big sample sizes and Mantel Test with small sample sizes.

In cases where the reference and focal group distributions differ, Mantel Test showed better performance in controlling the Type I error compared to IRT-LR. Thus, researchers or practitioners may prefer using Mantel Test primarily after evaluating the deviations in groups' ability distributions.

IRT-LR did not show good performance in controlling the Type I error when the amount of deviation increased for the focal group's ability distribution average and the sample increased. For this kind of test conditions, other DIF detection techniques (non-parametric DIF detection techniques) that are less sensitive to the mentioned test conditions should be used if this test will be used. Additionally, the results of these other DIF detection techniques should be evaluated with the results of this test and the researchers should decide whether or not the items display DIF or not taking all the results into consideration.

Acknowledgements

This paper is a part of unpublished doctoral dissertion Comparision of Mantel Test and Likelihood Ratio Test for Detection of Differential Item Functioning in Polytomous Item Responses

References

[1] S. Messick, "Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning", American Psychologist, 50 (9), pp. 741-749, 1995. doi: 10.1037/0003-066X.50.9.741

[2] Crocker, L and J. Algina, Introduction to Classical & Modern Test Theory. CA: Wadsworth Group, 1986, pp.217-236.

[3] W. H. Angoff, Perspectives on Differential Item Functioning Methodology. In P. W. Holland, and H. Wainer (Eds.), Differential Item Functioning (pp. 3-23). Hillsdale, NJ: Erlbaum, 1993.

[4] M. J. Gierl, J Bisanz, G. L. Bisanz, K. A. Boughton, and S. Khaliq,"Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests", Educational Measurement: Issues and Practice, vol. 20, pp. 26–36, 2001. doi: 10.1111/j.1745-3992.2001.tb00060.x

[5] C. F. Furlow, R. Fouladi, P. Gagné, and Whittaker, T. "A Monte Carlo study of the impact of missing data and differential item functioning on theta estimates under two polytomous Rasch family models", Journal of Applied Measurement, vol. 8 (4), pp. 388-403, 2007.

[6] R. D. Ankenmann, E. A Witt, and S. B. Dunbar, "An investigation of the power of the likelihood ratio goodness of fit statistic in detecting differential item functioning", Journal of Educational Measurement, vol. 36 (4), pp. 277-300, 1999. doi: 10.1111/j.17453984.1999.tb00558.x

[7] R. Zwick, J. R. Donoghue, and A. Grima, "Assessment of differential item functioning for performance tasks". Journal of Educational Measurement, vol. 30, pp. 233-251, 1993. doi:10.1111/j.1745-3984.1993.tb00425.x

[8] D. Thissen, L. Steinberg, and H. Wainer, "Detection of differential item functioning using the parameters of item response models", In P. W. Holland and H. Wainer (Eds.), Differential item functioning (pp. 67-113). Hillsdale NJ: Erlbaum, 1993.

[9] R. Ostini, and M. L. Nering, Polytomous Item Response Theory Models. CA: Sage, 2006.

[10] W.-C. Wang, and Y.-H Su, "Factors influencing the Mantel and generalized Mantel Haenszel methods for the assessment of differential item functioning in polytomous items", Applied Psychological Measurement, vol. 28, pp. 450-481, 2004. doi:10.1177/0146621604269792.

[11] D. Thissen, and H. Wainer, Test Scoring, New Jersey: Lawrence Erlbaum, 2001.

[12] J. P. Meyer, H Huynh, and M. A. Seaman, "Exact small-sample differential item functioning methods for polytomous items with illustration based on an attitude survey", Journal of Educational Measurement, vol. 41 (4), pp. 331-344, 2004. doi:10.1111/j.1745 3984.2004.tb01169.x

[13] P. Garrett, A Monte Carlo Study Investigating Missing Data, Differential Item Functioning and Effect Size. Unpublished doctoral dissertation, Georgia StateUniversity, 2009.

[14] H. Dodeen, "Stability of differential item functioning over a single population in survey data", Journal of Experimental Education, vol.72, pp. 181-193, 2004. doi: 10.3200/JEXE.72.3.181-193.

[15] W. S. Wood, "DIF testing for ordinal items with Poly-SIBTEST, the Mantel and GMH Tests, and IRT-LR-DIF when the latent distribution is non normal for both groups", Applied Psychological Measurement, vol. 35 (2), pp. 145–164, 2011. doi: 10.1177/0146621610377450

[16] B. Artar, Differential ItemFunctioning Analyses For Mixed Response Data Using IRT Likelihood-RatioTest, Logistic Regression and Gllamm Procedures. Unpublished doctoral dissertation, Florida State University, 2007.

[17] D. M. Bolt, "A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods", Applied Measurement in Education, vol. 15, pp. 113-141, 2002. doi:10.1207/S15324818AME1502_01

[18] K., A. Johnson- Frotman, the Evaluation of New Criteria for Polytomous DIF in the DFIT Framework. Unpublished doctoral dissertation. Illinois Institute of Technology, Chigago, 2007.

[19] J. V. Bradley, Robustness? British Journal of Mathematical & Statistical Psychology, vol. 31, pp144-152, 1978.

[20] Y. Chang, W. Huang, and R. Tsai, "DIF detection using multiple-group categorical CFA with minimum free baseline approach", Journal of Educational Measurement, vol. 52 (2), pp. 181-199, 2015. doi: 10.1111/jedm.12073

[21] P. Elosua, and C. Wells, "Detecting DIF in polytomous items using MACS, IRT and ordinal logistic regression", Psicologica: International Journal of Methodology and Experimental Psychology, vol. 34 (2), pp. 327-342, 2013.

[22] M. L. Bahry, Polytomous Item Response Theory Parameter Recovery: An Investigation of Non-Normal DistributionsAnd Small Sample Size. Unpublished master's thesis, University of Alberta, Canada 2012.