iojes
www.iojes.net

International Online Journal of Educational Sciences

ISSN: 1309-2707

Revisiting the Effect of Item Purification on Differential Item Functioning; Real Data Findings

Research Article

Emine Burcu TUNC[1], Muge ULUMAN[2], Akif AVCU[3]

[1]*Marmara University, Atatürk Faculty of Education, Istanbul, Turkey,* ORCID: 0000-0002-8225-9299

[2]*Marmara University, Atatürk Faculty of Education, Istanbul, Turkey,* ORCID: 0000-0003-4155-3114

[3]*Marmara University, Atatürk Faculty of Education, Istanbul, Turkey,* ORCID: 0000-0003-1977-7592

To cite this article: Tunc, E. B., Uluman, M., Avcu, A. (2018). Revisiting the Effect of Item Purification on Differantial Item Functioning; Real Data Findings, *International Online Journal of Educational Sciences*, 10(5), 139-147.

ARTICLE INFO

Article History:
Received: 11.06.2018
Available online: 13.11.2018

ABSTRACT

One of the important issues facing practitioners in the process of determining Differential Item Functioning (DIF) in a test, is that the presence of one or more items with DIF may affect the results of determining DIF in other items. In this case, items that do not function differentially could be falsely identified as showing DIF, which leads to an undesirable increase in Type I error. As a solution to this problem, it has been proposed that the items showing DIF are iteratively excluded from the analyses, in a process called item purification. The purpose of this study is to compare the results of gender based DIF analyses when item purification is conducted and when it is not. The study group consisted of 655 students who take undergraduate course of Measurement and Evaluation at a state university in Istanbul. The data collection tool consisted of 25 multiple choice items covering the curriculum in the course. Data analyses were performed using the R statistical program. The "difR" package was used for these analyses. The Mantel-Haenszel, Standardization, Logistic Regression, Lord's Chi-Square, Raju and Breslow Day methods were used during the purification and non purification processes. The findings showed that the DIF results with and without the iterative processes were changed and the numbers of DIF items showed difference.

Keywords:
Differential item functioning, iterative item purification, multiple-choice test

[1]Corresponding author's address: Marmara University, Atatürk Faculty of Education Goztepe Campus 34722 / Kadıköy - Istanbul
Telephone: 05063696073, e-mail: burcupehlivantunc@gmail.com
DOI: https://doi.org/10.15345/iojes.2018.05.010

INTRODUCTION

Differential Item Functioning (DIF) is a statistical measure of matching individuals according to their abilities in terms of the variable to be measured, and then providing evidence for those individuals in different groups whether they have different probabilities for responding to an item. (Camilli & Shepard, 1994; Zumbo, 1999). One of the major problems facing practitioners in the process of identifying DIF is that the presence of one or more items with DIF may affect the results of DIF detection in other items. For this reason, some items that do not function differently could be identified as showing DIF, which leads to an undesirable increase in Type I error (Magis et al., 2010). Type I error in the context of DIF identification, is the fact that when an item does not function differentially across the focal and reference groups, statistical analyses erroneously show that item to have DIF (Kim, 2010; Wyse & Mapuranga, 2009). Those items are called anchor items or DIF-free items (Magis et al., 2010). As a solution to this situation, it has been suggested that items displaying DIF should be iteratively excluded from the analyses in a process called item purification (Candell & Drasgow, 1988; Clauser, Mazor & Hambleton, 1993; Fidalgo, Mellenbergh & Muniz,2000; Holland & Thayer, 1988; Lautenschlager & Park, 1988; Wang & Su, 2004; Wang & Yeh, 2003).

Item purification has been discussed within the context of both Item Response Theory (Candell & Drasgow, 1988; Lautenschlager, Flaherty & Park, 1994; Park & Lautenschlager, 1990) and Classical Test Theory (Clauser, Mazor & Hambleton, 1993; French & Maller, 2007; Hidalgo-Montesinos & Gomez-Benito, 2003; Holland & Thayer, 1988; Miller & Oshima, 1992; Navas-Ara & Gomez-Benito, 2002; Wang & Su, 2004a, 2004b, 2010).

During the DIF detection process, comparable groups (e.g. women and men) are statistically matched before the responses are evaluated, according to the variables of interest (e.g. mathematics success). Matching has two types, internal and external. The observed score of the test is used with internal matching. The observed score of another test is used with external matching (Karami & Nodoushan, 2011). Matching that includes items showing DIF causes internal criteria to have inaccurate estimates of ability, and therefore incorrect DIF detection (Clauser, Mazor & Hambleton, 1993; Kim & Cohen, 1992). For this reason, it is very important to use purification processes for the matching criterion. The goal in purifying the matching criterion is to remove the DIF items determined in the initial analysis when calculating the matching criterion score (test total score) (Clauser & Mazor, 1998; French & Maller, 2007; Holland & Thayer, 1988).

There are two approaches for purification in the matching criterion. One of these is a two-stage approach proposed by Holland and Thayer (1988). The other is an iterative approach. The difference between the two approaches is the number of preliminary DIF analyses made to remove the DIF-detected items. If only one initial analysis is performed before removing the DIF showing items, these is called as two-step approach (Holland & Thayer, 1988). If the initial analyses is repeated until no DIF showing items remain, these is called iterative approach (French & Maller, 2007).

Fidalgo, Mellenbergh, and Muniz (2000) investigated conditions where two-stage purification, iterative purification, and no purification were performed during the DIF detection process and found that iterative purification performed better. At the same time, they also stated that iterative purification yielded better results when the number of DIF items was 15% and 30%.

Wang and Su (2004) also stated that two stage purification and iterative purification gave better results than situations where no purification was performed. The current study compared the results of iterative purification and no purification conditions in the DIF detection process by using data obtained from real testing conditions.

The basic principles of the method for the iteratively removing items are expressed in the following steps (Magis et al., 2010).

1. Testing all the items individually, assuming no item Show DIF.
2. Identification of DIF items according to the results of the step 1.
3. If there are no DIF items after the first iteration, step 6 is taken. Otherwise, continue with step 4.
4. Testing all of the items one by one by removing items identified as DIF at the end of the second step.
5. Identification of DIF items based on the results of step 4 and return of the third step.
6. Ending the process.

The statistical power in the context of DIF actually means that when an item function differently for the groups, that is, item shows DIF, to what extend the existence of DIF will be confirmed by the result of the statistical analyzes (Kim, 2010; Wyse & Mapuranga, 2009). It has been shown that the purification of items increases the statistical power in determining DIF (Fidalgo, Mellenbergh & Muniz, 2000; Lee & Geisinger, 2016). However, there are studies in which the opposite results are obtained.

French and Maller (2007) studied the effect of item purification using the LR method and indicated that iterarive purification does not contribute significantly to the increase of statistical power and to the control of Type I error. Magis and Facon (2012) investigated whether iterative item purification affected the amount of statistical power in small sample sizes was the delta plot method results showed that it did not affect it.

The studies that have investigated item purification during the DIF process generally use simulated data. In these studies, DIF and item purification were investigated under different simulation conditions. Lee and Geisinger (2016) found that when examining DIF and purification, the increase in statistical power for moderately difficult and highly discirimination power is greater than the increase in difficult and easy items. That is to say, items with high psychometric quality profit more by purification proces.

However, they found that statistical power increases when the number of DIF detected items is higher. Clauser et al. (1993) and French and Maller (2007) reached the same conclusions, but stated that the amount of Type I error is very low. Clauser et al. (1993), Wang and Su (2004) showed that, for different test lengths, the effect of item purification was different when the levels of ability between focal and reference groups differed. Also, when the test was short and there was a difference in ability between the two groups, they noted that item purification increased the Type I error. Fidalgo et al. (2000) stated that the number of DIF detected items may also affect item purification

Most of the studies that have been done so far are simulation studies. In this research, it is aimed to determine the effect of item purification on DIF results using real data collected in an actual testing situation.

METHODOLOGY

The Participants

The participants of this research is consisted of 655 students who took a undergraduate level course in Measurement and Evaluation at a stated U niversity located in Istanbul. Of the study group, 34.66% (227) were male and 65.34% (428) were female students. DIF analyses were performed according to gender.

Data Collection Tool

The data collection tool consisted of 25 items, prepared in a five-choice, multiple-choice format. The test was given as a final exam at the end of the semester and covered the curriculum of the course.

Analysis of the Data

Analyses of the data were performed using the R statistic (R Development Core Team, 2008). The "difR" package developed by Magis et al. (2010) was used for this purpose. In this study, Mantel-Haenszel, Standardization, Logistic Regression, Lord's Chi -Square, Raju and Breslow Day methods were used. Some theoretical information about the methods are given below.

Mantel-Haenszel (MH) Method

This method, developed by Mantel and Haenszel (1959), is suitable for testing the independence between two, two-categorical variables using a pair of K-partitioned data. For this reason, this method is based on the analysis of the *2x2xK* probability table.

Table 1. Mantel-Haenszel Table

Group	**1**	**0**	**Total**
Reference	A_j	B_j	N_{rj}
Focal	C_j	D_j	N_{rj}
Total	M_{1j}	M_{0j}	T_1,

Table 1 A_j and B_j correspond respectively to the numbers of correct and incorrect answers for any item. Similarly, C_j and D_j correspont to the numbers of correct and incorrect answers in the focal group. Using these obtained values, it is possible to calculate the MH statistic as follows:

$$MH\chi^2 = \frac{\left[\sum_J A_j - \sum_j E(A_j) - .5\right]^2}{\sum_j Var(A_j)}$$

In which, E(A_j) and Var (A_j) are expected value and variance of A_j. The -.5 correction in this equation is the continuity correction factor used to improve the estimation of the chi-square distribution, which is necessary for small frequencies. MH statistically monitors the chi-square distribution with a degree of freedom and examines whether there is a relationship between item response andgroup membership.

Standardization

In this method, developed by Dorans and Kulick (1986), the ratios of correct answers and each test score are compared in each group. The standard p-difference (ST-p-DIF) is the resulting test statistic and can be seen as the weighted averages of the differences in success rates between focal and reference groups (each level of the test score). Accordingly, the ST-p-DIF statistic takes the following form:

$$\text{ST-}p\text{-DIF} = \frac{\sum_j \omega_j(p_{fj} - p_{rj})}{\sum_j \omega_j}$$

Where P_{fj} and P_{rl} correspont to success rates in the focal and reference groups, respectively, and ω_j corresponds to the weighting system.

Logistic Regression Method

Developed by Swaminathan and Rogers (1990), this method essentially aims to overcome the limitations of the MH method. This method, which is used to determine uniform and non-uniform DIF, is based on logistic regression analyses that scores belonging to two categorized scoring factors are dependent variable, total test scores, variable indicating group membership and total score and predictive value of group membership interaction. The logistic regression equation is:

$$Y = b_0 + b_1 tot + b_2 form + b_3 tot*form$$

In which, *tot* stands for total scale score for each subject and Y is a natural logarithm of the probability ratio. So, the equation is:

$$\text{In}\left[\frac{p_i}{1-p_i}\right] = b_0 + b_1 tot + b_2 grup + b_3 tot{*}grup$$

Here, *p* corresponds to the proportion of individuals supporting the item in the latent variable.

For evaluation, 2-degrees of freedom Chi-Square test could be used for both uniform and non-uniform DIF.

Lord's Chi-Square Method

In this method developed by Lord (1980), the chi-square statistic is calculated by using the inverse of the variance-covariance matrix obtained by using the item parameter differences and the variance-covariance matrix obtained for these differences. Subsequently, this value is compared with a critical threshold based on a predetermined significance level. Here, the degrees of freedom correspond to the number of parameters examined for each item. If the observed Chi-Squared exceeds the critical value, the null hypothesis is rejected and the result of DIF is reached. The generated difference vector is expressed as:

$$V = (\hat{a}_F - \hat{a}_R \, , \, \hat{b}_F - \hat{b}_R)$$

Where $\hat{a}$ and $\hat{b}$ are the vectors of item parameter estimates for both the reference and the focal groups. The test statistic is expressed as:

$$Q = VS^{-1}V$$

Where *S* is the variance-covariance matrix of the differences between the item parameters. *Q* follows a chi-square distribution with a degree of freedom equal to the number of estimated parameters (Lord, 1980).

Raju Method

In this method, developed by Raju (1988), the area (signed) between the item characteristic curves of the focal and reference groups are calculated. The corresponding Z statistic is based on the null hypothesis that the real field is zero. A common metric must be created before the test can be performed. Any IRT model can be used in this approach. However, an important limitation is that for each item, the chance parameters for the group in both groups are equal. The Z-statistics for a one-parameter model can be simply calculated as:

$$Z = \frac{b_{jR} - b_{jF}}{\sqrt{\sigma_{jR}^2 + \sigma_{jF}^2}}$$

In which B_{jR} and σ_{jR} are are vectors of item parameter estimates and estimated standard errors respectively.

Breslow-Day Method (BD)

In this method, developed by Breslow and Day (1980), it is aimed to investigate whether the relationship between item response and group membership is homogeneous in the total test score range. Uniformity of DIF is present if homogeneity is not present. With the same notation as the MH method, the BD statistic can be written as:

$$BD = \sum_{j} \frac{\left[A_j - E(A_j)\right]^2}{Var(A_j)}$$

Further details for this formulation could be found in Aguerri, Galibert, Attorresi, & Marañón, 2009) and will not be given because of detailed derivations.

FINDINGS

Iterative material purification has been performed for each method used. The results obtained are presented in Table 2.

Table 2. Investigation of Purification Effect in Different DIF Methods

	M-H		**ST-p-DIF**		**L-R**		**Lord**		**Raju**		**B-D**			
	Purification													
	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	*f(X)*	*f(✓)*
Q1	-	-	-	-	-	-	-	-	-	-	-	-	0	0
Q2	-	-	-	-	-	-	-	-	-	-	-	-	0	0
Q3	-	✓	-	-	-	-	-	✓	-	✓	-	-	0	3
Q4	-	-	-	-	-	-	-	-	-	-	-	-	0	0
Q5	✓	✓	-	-	✓	✓	✓	✓	✓	✓	-	-	4	4
Q6	-	-	-	-	-	-	-	-	-	-	-	-	0	0
Q7	-	-	-	-	-	-	-	-	-	-	-	-	0	0
Q8	-	✓	-	-	-	-	✓	✓	✓	✓	-	-	2	3
Q9	-	-	-	-	-	-	-	-	-	-	-	-	0	0
Q10	-	-	-	-	-	-	-	-	-	-	-	-	0	0
Q11	-	-	-	-	-	-	-	-	-	-	-	-	0	0
Q12	-	-	-	-	-	-	-	-	-	-	-	-	0	0
Q13	✓	-	-	-	-	-	-	-	-	-	-	-	1	0
Q14	-	-	-	-	-	-	-	-	-	-	✓	✓	1	1
Q15	-	-	-	-	-	-	-	-	-	-	-	-	0	0
Q16	-	-	-	-	-	-	-	-	-	-	-	-	0	0
Q17	-	-	-	-	-	-	-	-	-	-	-	-	0	0
Q18	-	-	-	-	-	-	-	-	-	-	-	-	0	0
Q19	-	-	-	-	✓	✓	-	-	-	-	-	-	1	1
Q20	-	-	-	-	-	-	-	-	-	-	-	-	0	0
Q21	-	-	-	-	-	-	-	-	-	-	-	-	0	0
Q22	-	-	-	-	-	-	-	-	-	-	-	-	0	0
Q23	-	-	-	-	-	-	-	-	-	-	-	-	0	0
Q24	-	-	-	-	-	-	-	✓	-	✓	-	-	0	2
Q25	-	-	-	-	-	-	-	-	-	-	-	-	0	0
f (method.)	2	3	0	0	2	2	2	4	2	4	1	1		

Table 2 contains the findings of six different DIF detection methods with and without purification. First, each method was examined in terms of the number of DIF items observed for both purification processes. In this context, there are three methods (ST-*p*-DIF, LR and BD) in which the number of items showing DIF is the same regardless of the purification conducted. Three methods (MH, Lord and Raju) produce fewer DIF items

when the purification process is not performed. In the absence of purification, there is no method showing more DIF items.

When each method was examined in detail, the Std. method did not show any items with DIF independently of the purification process. The methods L-R (5th and 19th items) and BD (14th item) detected different items as showing DIF similarly when and without purification. When the results for the MH method were examined, it was found that two items (5th and 13th) were detected when purification was not applied while three items showed DIF (3th, 5th and 8th) after the purification process used. When the MH method was examined, it showed two DIF items (5 and 13) with no purification process and it showed three DIF items (3, 5 and 8) after the purification process was used. The Lord and Raju methods showed the same results with and without the purification process. According to both methods, the 5th and 8th items appeared to have DIF for both conditions, and in addition to these, two items (4th and 24th ones) showed DIF when purification was applied.

According to the findings obtained, Lord and Raju methods are the most sensitive ones for purification process are. Two items were identified as showing DIF without purification, while four items emerged as with DIF when purification carried out. The methods that are the same accross two conditions are LR, ST-*p*-DIF and BD methods.

When the findings in Table 2 are examined in terms of items, it is noteworthy that the 5th item was detected as showing DIF regardless of the purification process, in four of the six methods. In addition, 8th item is one the most captured by the methods as showing DIF after the 5th item. Based on two methods (Lord and Raju), this item was identified as DIF showing in both purification conditions. In addition, according to based on one (MH) method, DIF was captured when purification is carried out.

Third item was found to show DIF when purification used in three different methods (M-H, Lord and Raju). Twenty fourth item appears as showing DIF when the purification carried out in the two methods while Item 14 and item 19 appear as showing DIF independently purification process based on the findings from only one method (BD, LR, respectively). Finally, 13th item is marked as showing DIF only when purification was not applied in one method (M-H).

DISCUSSION

In this study, it has been revealed that the purification process affects the number of items detected as showing DIF. Research has also suggested that purification during the DIF detection affects the results and that purification should be used during this process (Candell & Drasgow, 1988, Clauser, Mazor & Hambleton, 1993, Fidalgo, Mellenbergh & Muniz, 2000, Holland & Thayer, 1988; Lautenschlager & Park, 1988; Wang & Su, 2004; Wang & Yeh, 2003). In this context, it can be stated that the findings reached correspond to the literature.

According to the findings, the most sensitive method to the purification process is Lord's Chi-Square and Raju methods. It is point out that these methods are IRT based methods. In this direction, it is possible to reach to the conclusion that the results of purification process can affect IRT-based methods more.

Simulated data are generally used in studies examining the effect of purification on DIF. Most of these studies focused on the amount of change to statistical power and Type I error (Candell & Drasgow, 1988, Clauser, Mazor & Hambleton, 1993, French & Maller, 2007, Hidalgo-Montesinos & Gomez-Benito, 2003, Holland & Thayer, 1988; Lautenschlager, Flaherty & Park, 1994; Navas-Ara & Gomez-Benito, 2002; Park & Lautenschlager, 1990; Wang & Su, 2004a, 2004b, 2010). These studies also found that the purification process affected the statistical power and Type I error rates. Findings of this study were obtained from real data. Therefore, statistical power and Type I error were not discussed here. However, findings of the current study were similar to these simulation studies and the effect that item purification have an effect on DIF results.

REFERENCES

Aguerri, M. E., Galibert, M. S., Attorresi, H. F., & Marañón, P. P. (2009). Erroneous detection of nonuniform DIF using the Breslow–Day test in a short test. *Quality & Quantity*, 43, 35-44.

Breslow, N. E., & Day, N. E. (1980). *Statistical methods in cancer research: Vol. 1. The analysis of case–control studies* (Scientific Publication No. 32). Lyon, France: International Agency for Research on Cancer.

Camilli G. and Shepard L. A. (1994). *Methods for identifying biased test items (volume 4).* California: SAGE Publications. Inc.

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement*, 12, 253-260.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differential functioning test items. *Educational Measurement: Issues and Practice*, 17(1), 31-44.

Clauser, B. E., Mazor, K. M., & Hambleton, R. K. (1993). The effects of purification of the matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education*, 6, 269-279.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning. Standardization and the Mantel–Haenszel method. *Applied Measurement in Education*, 2, 217-233.

Fidalgo, A. M., Mellenbergh, G. J., & Muniz, J. (2000). Effects of amount of DIF, test length and purification type on robustness and power of Mantel-Haenszel procedures. *Methods of Psychological Research Online,* 5, 43-53.

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for differential item functioning detection. *Educational and Psychological Measurement* 67, 373-393.

Hidalgo-Montesinos, M.D., & Gómez-Benito, J. (2003). Test purification and the evaluation of differential item functioning with multinomial logistic regression. *European Journal of Psychological Assessment*, 19, 1-11.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel- Haenszel procedure. In H. Wainer & H. Braun (eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Erlbaum.

Karami H. and Nodoushan M. A. S. (2011). Differential item functioning (DIF): current problems and future directions. *International Journal of Language Studies*, 5-4, 133-142.

Kim, J. (2010). *Controlling type 1 error rate in evaluating differential item functioning for four DIF methods: Use of three procedures for adjustment of multiple item testing* (Doctoral Dissertation). Georgia State University, ABD.

Lautenschlager, G. J., Flaherty, V. L., & Park, D. G. (1994). IRT differential item functioning: An examination of ability scale purifications. *Educational and Psychological Measurement*, 54, 21-31.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Magis, D., & Facon, B. (2012). Item purification does not always improve DIF detection: A counterexample with Angoff's delta plot. *Educational and Psychological Measurement*, 73, 293-311.

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior research methods*, 42(3), 847-862.

Miller, M. D., & Oshima, T. C. (1992). Effect of sample size, number of biased items, and magnitude of bias on a two-stage item bias estimation method. *Applied Psychological Measurement*, 16, 381-388.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719-748.

Navas-Ara, M. J., & Gómez-Benito, J. (2002). Effects of ability scale purification on identification of DIF. *European Journal of Psychological Assessment*, 18, 9-15.

Park, D. G., & Lautenschlager, G. J. (1990). Improving IRT item bias detection with iterative linking and ability scale purification. *Applied Psychological Measurement*, 14, 163-173.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, 53, 495-502.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27, 361-370.

Wang, W.-C., & Su, Y.-H. (2004). Effects of average signed area between two item characteristic curves and test purification procedures on the DIF detection via the Mantel-Haenszel method. *Applied Measurement in Education*, 17, 113-144.

Wang, W.-C., & Yeh, Y.-L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement*, 27, 479-498.

Wang,W.-C., & Su, Y.-H. (2004). Factors Influencing the Mantel and generalized Mantel- Haenszel methods for the assessment of differential item functioning in polytomous items. *Applied Psychological Measurement*, 28, 450-480.

Wang,W.-C., & Su, Y.-H. (2010). MIMIC Methods for Assessing Differential Item Functioning in Polytomous Items. *Applied Psychological Measurement*, 34, 166-180.

Wyse, A. E. and Mapuranga, R. (2009). Differential item functioning analysis using Rasch item information functions. *International Journal of Testing*, 9(4), 333-357.

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.