Available online at www.elixirpublishers.com (Elixir International Journal)

Psychology

Elixir Psychology 110 (2017) 48252-48258

ISSN: 2229-712X

Effect of sample size, ability distribution and test length on detection of differential item functioning using logistic regression statistic

Ferdinand Ukanda[1], Lucas Othuon[1], John Agak[1] and Paul Oleche[2]

[1]Department of Educational Psychology, Maseno University, Private Bag, Maseno, Kenya.

[2]Department of Pure and Applied Mathematics, Maseno University, Private Bag, Maseno, Kenya.

ARTICLE INFO

Article history:

Received: 17 July 2017;

Received in revised form:

2 September 2017;

Accepted: 14 September 2017;

Keywords

Differential Item Functioning (DIF),

Logistic Regression (LR),

Effect size (ES),

Sample size,

Ability distribution,

Test length,

WinGen3.

ABSTRACT

Differential Item Functioning (DIF) is a statistical method that determines if test measurements distinguish abilities by comparing two sub-population outcomes on an item. The Logistic Regression (LR) statistic provides an effect size measure that can give the magnitude of DIF. The purpose of the study was to investigate through simulation the effects of sample size, ability distribution and test length on the Effect Size (ES) of DIF and their influence on detection of DIF using LR method. A Factorial research design was used in the study. The population of the study consisted of 2000 examinee responses. A stratified random sampling technique was used with the stratifying criteria as the reference (r) and focal (f) groups. A small sample size (60r/60f) and a large sample size (1000r/1000f) were established. WinGen3 statistical software was used to generate dichotomous item response data which was replicated 1000 times. The findings of the study showed that whereas sample size and ability distribution had significant effects on the ES of DIF items when LR was used, test length had no statistically significant effect on the ES of DIF items. However, the number of DIF detections using LR statistic increased with test length regardless of the nature of Ability Distribution, The findings of the study are of great significance to teachers, educational policy makers, test developers and test users.

Introduction

Background to the Study

Differential item functioning (DIF) analysis is typically used to identify test items that are differentially difficult for respondents who have the same ability level of knowledge or skill but differ in ways that should be irrelevant to their performance on a test. DIF is a collection of statistical methods utilized to determine if examination items are appropriate and fair for testing the knowledge of different groups of examinees. DIF methods therefore assess the test-takers' response patterns to specific test items. Conclusions drawn about group differences among examinee groups should therefore be accurate. The accuracy of a DIF detection statistic can be determined by the magnitude of the effect size measure under different conditions. Several Monte Carlo DIF detection studies have focused on the influence of sample size on DIF detection to determine the sample size that results in minimal variance and least error rates with DIF detection procedures (Gonzalez & Roma, 2006).

The Logistic Regression (LR) procedure is one of the most common procedures for detecting differential item functioning (Wang & Su, 2004; Swaminathan & Rodgers, 1990). Jodoin and Gierl (2002) showed that test length had no significant influence on the power of the LR procedure for DIF detection. Uttaro and Millsap (1994) used both short (20 items) and moderate (40 items) test lengths, but DIF was presented only in the studied item. Test length generally had little effect on the detection rates in both the 20- and 40 item tests. DIF methodology also assumes that ability distribution for the focal and reference groups are equal. In this simulation study, the ability distribution for the focal and reference groups was varied.

A study by Pedrajita and Talisayon (2009) identified biased test items through differential item functioning analysis using Logistic Regression. The study made use of test scores of 200 junior high school students. One hundred students came from a public school, and the other 100 were private school examinees. One hundred students were males and 100 were females. Basing from their English II grades, 95 students were classified as low ability and 105 as high ability students. A researcher-constructed and validated Chemistry Achievement Test was used as research instrument. The results from the method used were compared, and it was found that school type, gender, and English ability bias existed. Logistic Regression Statistic identified biased test items.

The Logistic Regression (LR) method has been one of the common methods in DIF research (Wang & Su, 2004; Swaminathan & Rogers, 1990). The method is currently seen as a practical means of determining DIF because of its simplicity and ease of use, and providing an effect size statistic to determine if the DIF found is damaging.

It uses the item response (0 or 1) as the dependent variable, with grouping variable (dummy coded as 1= reference, 2=focal), total scale score for each subject (characterized as variable TOT) and a group by TOT interaction as independent variables. This method provides a test of DIF conditionally on the relationship

Tele:
E-mail address: ukanda2001@yahoo.com

between the item response and the total test score, testing the effects of group for uniform DIF, and the interaction of group and TOT to assess non-uniform DIF. Uniform DIF exists when there is no interaction between ability level and group membership. The presence of DIF in the LR approach is determined by testing the improvement in model fit that occurs when a term for group membership and a term for the interaction between test score and group membership are successively added to the regression model. A chi-square test is then used to evaluate the presence of uniform and non-uniform DIF on the item of interest by testing each term included in the model. The general model for Logistic Regression takes the form:

$$\boldsymbol{p(u=1)=\frac{e^{z}}{1+e^{z}}}$$

where u is the score on the studied item. Performance on the studied item is first conditioned on the total test score. In this step, z = β0 + β1 X where X is the test score (Model 1). This serves as the baseline model. The presence of uniform DIF is then tested by examining the improvement in chi-square model fit associated with adding a term for group membership (G) against the baseline model. That is, Model 2 (i.e. z = β0 + β1 X + β2 G) subtracted from Model 1. The presence of no uniform DIF is tested by examining the improvement in chi-square model fit associated with adding a term for group membership (G) and a term for the interaction between test score and group membership (XG) against model 2. In other words, Model 3 (i.e. z = β0 + β1 X + β2 G + β3 XG) subtracted from Model 2. Zumbo and Thomas (1996) developed an index to quantify the magnitude of DIF for the LR procedure based on partitioning a weighted least-squares estimate of R2 that yields an effect size measure. This index is obtained, first, by computing the R2 measure of fit DIF for each term in the LR model (i.e., test score, group membership, test score-by-group membership interaction) and then by partitioning the R2 for each of the terms. A DIF effect size for the group membership term is produced by subtracting the R2 for the group membership term (Model 2) from the R2 for the total test score term (Model 1). The result is an effect size measure associated with group membership that quantifies the magnitude of uniform DIF (herein called R2Δ - U). A second DIF effect size is produced for the total score-by-group membership term by subtracting the R2 for` the group membership interaction that quantifies the magnitude of non-uniform DIF (herein called R2Δ - N). R2 Δ can be used with the LR significance test to identify items with DIF. Jodoin (1999) empirically-established guidelines for interpreting R2Δ. An item has negligible or A-level DIF when the chi-square test for model fit is not statistically significant or when R2Δ < 0.035. An item has moderate or B-level DIF when the chi square test is statistically significant and when 0.035 ≤ R2 Δ < 0.070. An item has large or C-level DIF when the chi-square test is statistically significant and when R2Δ ≥ 0.070. These guidelines are applicable to both uniform and non-uniform DIF, and were used to classify DIF items in the current study.

Objectives of the Study

The objectives of the study were to:

(i) Determine the effect of Sample Size, Ability Distribution and Test Length on the Effect Size of DIF items across 3 DIF Types; A, B and C using LR statistic.

(ii) Investigate the influence of Sample Size, Ability Distribution and Test Length on the number of detections of DIF items across 3 DIF Types; A, B and C using LR statistic.

Methodology

Research Design

A factorial research design was used in this study. This design was used to simulate samples for different conditions resulting into a 3 x 3 x 2 factorial design giving 18 data sets. The independent factors were sample size, type of ability distribution, and test length. The dependent factor was the number of DIF items detected based on the magnitude of the effect sizes.

Sample and Sampling Technique

A stratified random sampling technique was used to select the sample from a pool of 2000 examinee responses. The stratifying criterion was based on the examinee responses designated as reference and focal. The reference and focal groups had three sample sizes each namely: 20, 60, and 1000. These were used to establish three sample size conditions namely two small sample sizes [(20*r*/20*f*), (60*r*/60*f*)], and one large sample size (1000*r*/1000*f*).

Data Collection Procedure

WinGen3 (Han, 2009) statistical software was used to generate dichotomous item response data. The main window consisted of examinee characteristics which included the number of examinees and the ability distribution in terms of mean and standard deviation. It also consisted of item characteristics which included the number of items, the number of response categories, the model to be used i.e. 1PLM, 2PLM, 3PLM or non-parametric. The distribution in terms of parameter *a*, *b* and *c* was selected. When appropriate entries were made, true scores and true item parameters were then generated. Replication data sets and response data sets were also generated. The software allowed examinee graphs and item graphs to be displayed. The DIF/IPD window consisted of introduction to DIF/Item parameter drift via the direct input mode or the multiple file read in mode. This consisted of data files for the reference group/test 1 and focal group's later tests. Binary response data representing examinee responses on a test were generated. The user then chose typical test lengths to make the simulation data approximate real data as much as possible. The tests had 10 items, 30 items and 50 items respectively. The software was also used to vary the ability distribution of the data. The obtained data was replicated 1,000 times for every cell in the study, resulting into 18,000 data sets. The average value of the effect sizes across the 1000 replications was calculated.

Methods of Data Analysis

Analysis was done using the Statistical Package for Social Sciences (IBM SPSS Version 20) computer software. It used the General Linear model, multivariate analysis which gave R^2 values for model 1 and model 2. The R^2 values were then entered into coding sheets on MS Excel worksheet to obtain the Effect size, R^2 Δ which was the difference between R^2 values for model 1 and model 2. The procedure was repeated for 1000 replications and the average Effect size value was determined. The number of items displaying various categories of DIF were then determined for each category of Test length. One Way Analysis of Variance (ANOVA) was used to determine the effect of Sample Size, Ability Distribution and Test Length on the Effect Size (ES) of DIF and detection of DIF across three types of DIF; A, B and C.

Line graphs for mean Effect size against Test length Across DIF types and for each level of Ability distribution and Sample size were constructed to aid interpretation. A similar display for the mean number of items across various categories of DIF was constructed.

Results

Effect Size for Different Item Types under Different Conditions

The effect sizes for different types of DIF items under different conditions are presented in Table 1. As would be expected, the ES for Type A DIF items had the smallest values and those for Type C items had the largest values.

Table 1. Effect size for different types of DIF items under different conditions.

No. of items	Ability distribution (Mean, SD)	Sample size	Effect size		
			Type A	Type B	Type C
10	(0, 1)	20	0.02313	0.04440	0.28740
10	(1, 2)	20	0.02020	0.04343	0.21416
10	(0, 1)	60	0.02185	0.04270	0.17816
10	(1, 2)	60	0.01551	0.06333	0.28640
10	(0, 1)	1000	0.00783	0.05232	0.15490
10	(1, 2)	1000	0.00592	0.05500	0.17635
30	(0, 1)	20	0.02842	0.04803	0.14392
30	(1, 2)	20	0.02484	0.04406	0.19029
30	(0, 1)	60	0.01647	0.04652	0.13831
30	(1, 2)	60	0.01890	0.05111	0.21542
30	(0, 1)	1000	0.00999	0.04242	0.28019
30	(1, 2)	1000	0.01242	0.05753	0.27430
50	(0, 1)	20	0.02727	0.04878	0.22616
50	(1, 2)	20	0.02579	0.04738	0.20307
50	(0, 1)	60	0.01977	0.05089	0.18840
50	(1, 2)	60	0.01599	0.05474	0.35606
50	(0, 1)	1000	0.00793	0.04673	0.20589
50	(1, 2)	1000	0.00865	0.05390	0.25412

Key: Type A=Negligible DIF, Type B=Moderate DIF, Type C=Large DIF

Effect of Sample Size on Effect Size of DIF across DIF Types using LR Statistic

In order to determine the effect of Sample Size on effect size for each type of DIF items, one-way analysis of variance was conducted with Effect Size as the dependent variable and Sample Size as the independent variable.

Table 2. ANOVA results for the effect of Sample Size on the ES of DIF across 3 DIF Types using LR statistic.

Type of DIF		Sum of Squares	df	Mean Square	F	Sig.
A	Between Groups	.00078854	2	.00039427	59.2256	.000000076
	Within Groups	.00009986	15	.00000666		
	Total	.00088840	17			
B	Between Groups	.00083037	2	.00041518	1.36845	.28451104
	Within Groups	.00455094	15	.00030340		
	Total	.00538131	17			
C	Between Groups	.00090915	2	.00045457	.119424	.88826636
	Within Groups	.05709560	15	.00380637		
	Total	.05800475	17			

Statistically significant differences between means was recorded for the A Type of DIF only (F=59.2256, df_b=2, df_w=15, p=.000000076). Post-hoc analysis using Bonferroni method for pairwise comparisons revealed that for A Type DIF items, differences existed between sample size 20 and 60; and 20 and 1000; and 60 and 1000 only as displayed in Table 3.

Effect of Ability Distribution on Effect Size of DIF across DIF Types

In order to determine the effect of Ability Distribution on ES for each type of DIF items, one-way analysis of variance was conducted with ES as the dependent variable and Ability Distribution as the independent variable.

Effect of Test Length on Effect Size of DIF across 3 DIF Types

In order to determine the effect of Test Length on ES for each type of DIF items, one-way analysis of variance was conducted with ES as the dependent variable and Test Length as the independent variable. Table 5 summarizes the ANOVA results for the effect of Test Length on the ES of DIF across 3 DIF Types using LR statistic. The findings indicate that Test Length had no statistically significant effect on ES of DIF items regardless of the type of DIF (*p*>.05).

Table 3. Pairwise comparisons of effect sizes across different sample sizes for Type A DIF.

(I) Sample size	(J) Sample size	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
20	60	.0068600*	.00148964	.001	.0028473	.0108727
	1000	.0161517*	.00148964	.000	.0121390	.0201644
60	20	-.0068600*	.00148964	.001	-.0108727	-.0028473
	1000	.0092917*	.00148964	.000	. 0052790	.0133044
1000	20	-.0161517*	.00148964	.000	-.0201644	-.0121390
	60	.0092917*	.00148964	.000	-. 0133044	-.0052790

Table 4. ANOVA Summary for effect of Ability Distribution on effect size of DIF for LR across 3 DIF types.

Type of DIF		Sum of Squares	df	Mean Square	F	Sig.
A	Between Groups	.00001158	1	.00001158	.211385	.65186968
	Within Groups	.00087681	16	.00005480		
	Total	.00088840	17			
B	Between Groups	.000015272	1	.000015272	.045537	.83371535
	Within Groups	.005366038	16	.000335377		
	Total	.005381310	17			
C	Between Groups	.007476120	1	.007476120	2.36736	.14343733
	Within Groups	.050528549	16	.003158034		
	Total	.058004748	17			

There were no statistically significant differences for the effect of Ability Distribution on ES.

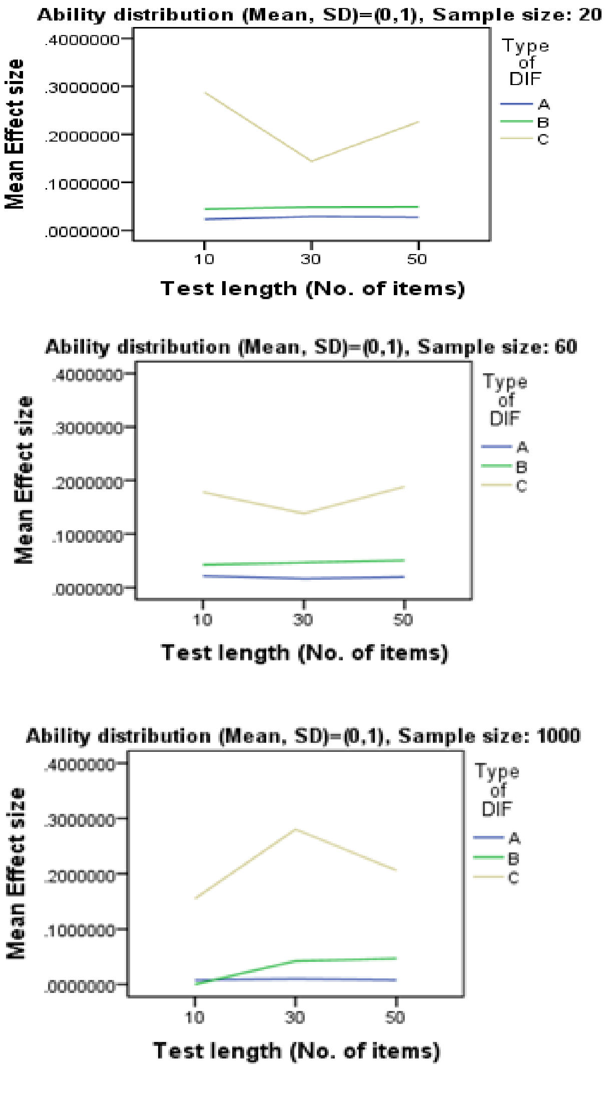
Table 5. ANOVA Summary for effect of Test Length on effect size of DIF for LR across 3 DIF types.

Type of DIF		Sum of Squares	df	Mean Square	F	Sig.
A	Between Groups	.00002375	2	.00001187	.206004	.81609397
	Within Groups	.00086465	15	.00005764		
	Total	.00088840	17			
B	Between Groups	.00078593	2	.00039297	1.28270	.30601893
	Within Groups	.00459380	15	.00030636		
	Total	.00538131	17			
C	Between Groups	.00323269	2	.00161635	.442656	.65045334
	Within Groups	.05477205	15	.00365147		
	Total	.05800475	17			

Further to the above analyses, line graphs were constructed for mean ES against Test Length across DIF types and for each level of Ability Distribution and Sample Size. This outcome is presented in Figure 1 to aid more detailed interpretation of data.

The largest mean ES was recorded for Type C DIF items. This was followed by Type B and A, respectively. This outcome was regardless of Ability Distribution, Sample Size and Test Length. However, differences in ES between A and B items were not as large as those between either A and C or B and C items.

ABILITY DISTRIBUTION WITH MEAN=0, SD=1



ABILITY DISTRIBUTION WITH MEAN=1,SD=2

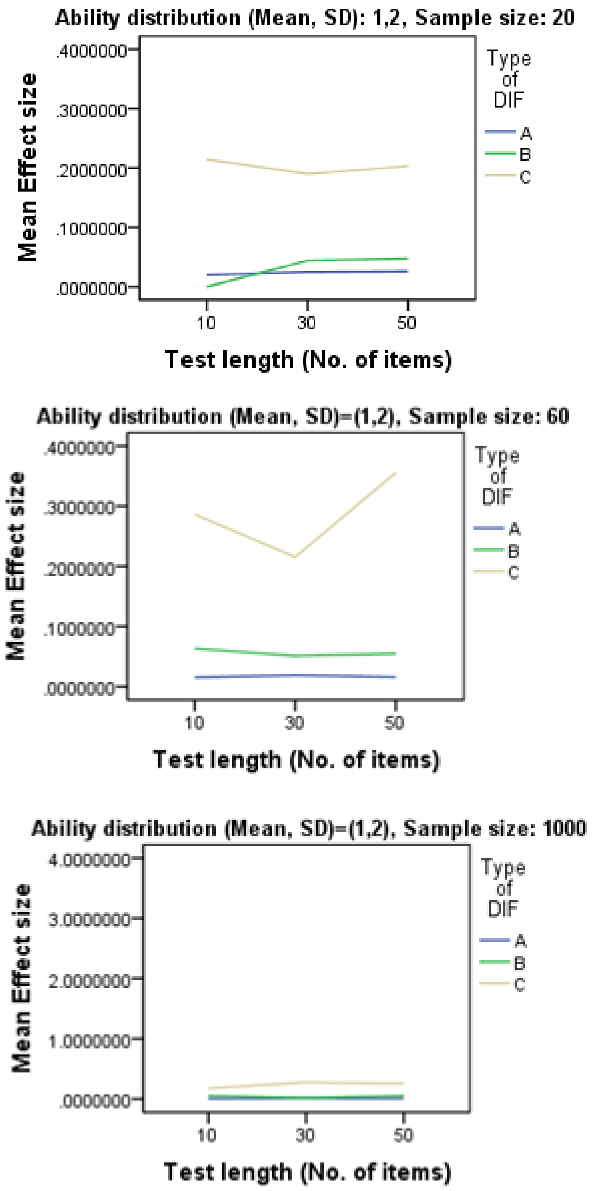


Figure 1: Mean effect sizes for different types of DIF under different conditions.

More specifically, for Ability Distribution with (Mean, SD)=(0, 1) and Sample Size=20, mean ES was largest for Type C items followed by B and A. However, the highest ES for Type C items occurred for 10 items. For Type C items, when Ability Distribution had (Mean, SD)=(1, 2) and Sample Size=20, the smallest ES was recorded at Test Length=30 items. For Ability Distribution with (Mean, SD)=(1, 2) and Sample Size=20, the mean ES was largest for Type C items followed by B and A. For Type C DIF items, the largest ES was recorded for 10 items and the smallest for 30 items with the magnitude of ES decreasing with Test Length. For Type B ES tended to marginally increase with Test Length while for type A it remained constant with an increase in test length.

For Ability Distribution with (Mean, SD)=(0, 1) and Sample Size=60, the mean ES was largest for Type C items followed by B and A. For Type C DIF items in this category, the largest ES was recorded for 50 items and the smallest for 30 items. For Type A and B, ES tended to remain constant with Test Length. This trend was reasonably maintained when the Ability Distribution with (Mean, SD)=(1, 2) and Sample Size=60 though the mean effect size for type C was larger tha that for (Mean, SD)=(0, 1) and Sample Size=60.

For Ability Distribution with (Mean, SD)=(0, 1) and sample size=1000, mean ES was largest for Type C items followed by B and A. The largest ES for Type C items in this category was recorded for 30 items and the smallest for 10 items. For Type C items, when Ability Distribution had (Mean, SD)=(1, 2) and Sample Size=1000, the largest ES was recorded at Test Length of 30 items. The mean effect size of Type C items was much lower than that when the Ability Distribution had (Mean, SD)=(0, 1) and Sample Size=1000. For Type A and B items the ES was very low but also tended to be the same across the various test lengths. Ability distribution therefore tended to have an effect on the ES regardless of the Sample size and Test length.

Number of DIF Items Detected under Different Conditions The number of DIF items detected under different conditions is shown in Table 6 for three types of DIF items; A, B and C. The information in Table 6 is summarized using line graphs in Figure 2. The graphs show the mean number of detections for different types of DIF under different conditions of Sample Size, Ability Distribution and Test length.

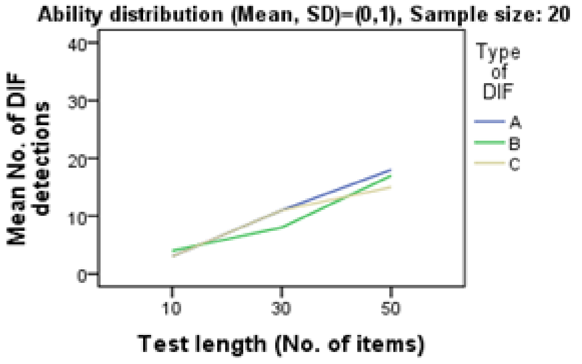
Table 6. Number of DIF items detected under different conditions.

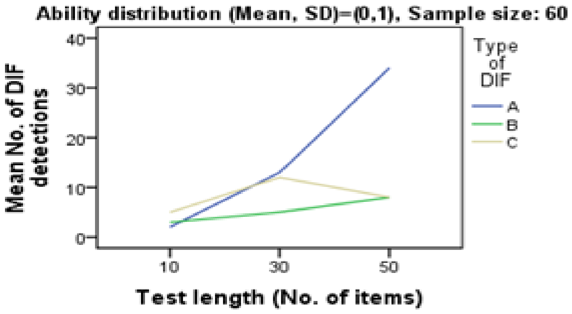
No. of items	Ability distribution (Mean, SD)	Sample size	Number of DIF detections		
			Type A	Type B	Type C
10	(0, 1)	20	3	4	3
10	(1, 2)	20	2	0	8
10	(0, 1)	60	2	3	5
10	(1, 2)	60	7	1	2
10	(0, 1)	1000	9	0	1
10	(1, 2)	1000	5	3	2
30	(0, 1)	20	11	8	11
30	(1, 2)	20	9	5	16
30	(0, 1)	60	13	5	12
30	(1, 2)	60	9	9	12
30	(0, 1)	1000	18	4	8
30	(1, 2)	1000	7	1	22
50	(0, 1)	20	18	17	15
50	(1, 2)	20	12	6	32
50	(0, 1)	60	34	8	8
50	(1, 2)	60	13	8	29
50	(0, 1)	1000	32	4	14
50	(1, 2)	1000	21	5	24

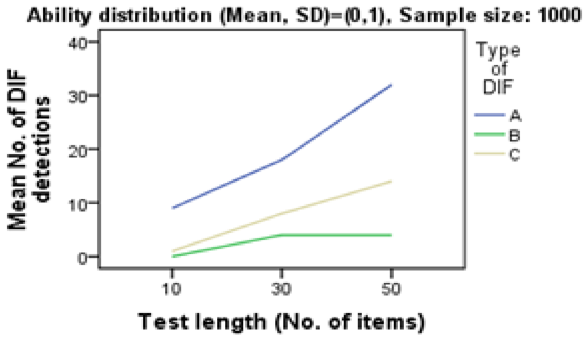
Key: Type A=Negligible DIF, Type B=Moderate DIF, Type C=Large DIF

In general, the mean number of DIF detections using LR statistic increased with Test Length regardless of the nature of Ability Distribution, Sample Size and Type of DIF. When the Ability Distribution was such that (Mean, SD)=(0, 1), and the

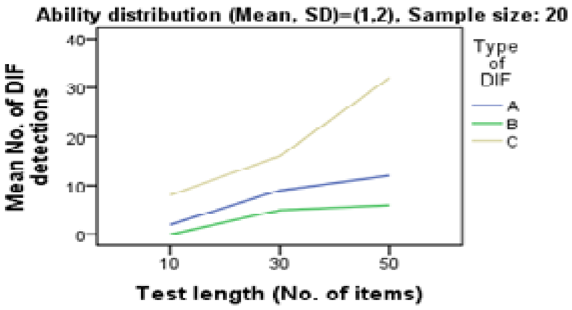
ABILITY DISTRIBUTION WITH MEAN=0,SD=1

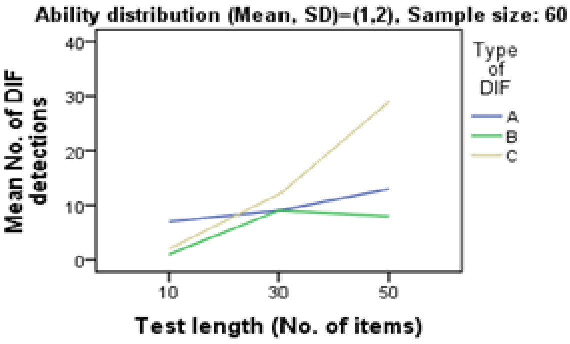






ABILITY DISTRIBUTION WITH MEAN=1, SD=2





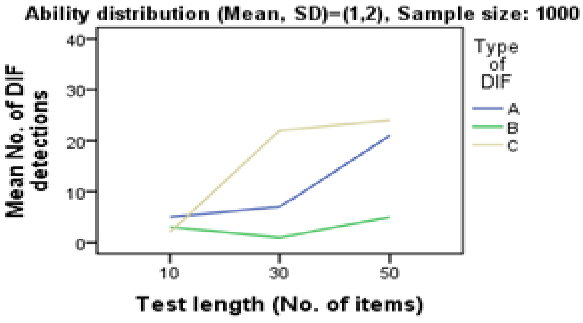


Figure 2. Mean number of DIF detections for different types of DIF under different conditions.

Sample Size was at its lowest level of 20, only marginal differences in DIF detection occurred between Type A and Type C items. However, there were reasonable differences in DIF detection between the two item types and Type B items, with the highest mean DIF detection being recorded for Type B items.

In addition, the largest difference in DIF detection was recorded when Test Length was 50 items (Large Test Length). The same pattern was maintained when Sample Size increased to 60 except that the DIF detection between Type A and Type B DIF items at this level tended to increase as Test Length increased to 30 and then to 50 items.

When Sample Size=1000 and Ability Distribution was (Mean, SD)=(0, 1), differences in mean DIF detection were large across the three types of DIF items i.e. A, B and C. However, differences in mean number of DIF detection tended to increase with Test Length, with the largest difference occurring when Test Length was 50 items i.e. for the longest test. A point of departure from the previous two trends is that in this case (i.e. Sample size of 1000 and Ability Distribution with (Mean, SD)=(0,1), Type A items were detected much more than Type C items for the case of the longest test with 50 items.

At Sample Size=20 and Ability Distribution with (Mean, SD)=(1, 2), Type C items consistently recorded the highest mean number of DIF detections across the three levels of test length (i.e. 10, 30 and 50 items). The smallest difference in mean number of DIF detections in this case was found to exist between Type A and Type B items for the shortest test of 10 items. For a sample of size 60, the difference in mean DIF detection for Type B and Type C items was minimal for a test length of 10 items but it was very large for a test length of 50 items. The same number of DIF items was recorded for Type A and B for a test length of 30 items. When sample size got increased to 1000, results were similar to those for sample size of 60 except that Type A and Type B items exhibited relatively larger differences in mean DIF detection at a test length of 50 items. Thus, when the ability distribution has (Mean, SD)=(0, 1),and number of items is large (32), LR statistic gives optimal results for Type A items than for Type B or C items.

Limitations of the Study

This study made use of dichotomous item response data and not polytomously scored items. It is important that care is taken not to generalize findings to polytomous data as this was outside the scope of the present study.

While the results reveal significant findings and draw important implications in the field of DIF, Harrison et al. (2007) argue that simulation is prone to misspecification errors. Further, Davies, Eisenhardt and Bingham (2007) also observed that generalization based on simulation studies must be treated with caution beyond the parameter range specified in the model. This notwithstanding, it is important to mention that Othuon (1998), and Davies, Eisenhardt and Bingham (2007) observed that the key strength of simulation is its ability to support investigation of phenomena that are hard to research by conventional means, particularly in situations where empirical data are limited.

Discussion

The purpose of this study was to investigate the effect of Sample Size, Ability Distribution and Test Length on Effect Size (ES) of DIF, and the influence of the same variables on detection of DIF using Logistic Regression (LR) statistic. Results indicate that Sample Size had a statistically significant effect on ES for A Type items (Negligible DIF items) and not for B or C Types. Post-hoc test indicated that significant differences in ES for A Type items existed between Sample size 20 and 60; and 20 and 1000; and 60 and 1000 only. This suggests that it is A Type items that may be problematic when measuring DIF using MH statistic, particularly for negligible to large sample sizes.

Ability Distribution was found to have a statistically significant effect on ES for C Type items (i.e. Large DIF items) only. This suggests that for items with large DIF, the nature of Ability Distribution remains crucial when using the LR statistic.

Test Length had a statistically significant effect on ES for Type B DIF item Types. There was a general trend for ES to increase with Test Length. This is inconsistent with the findings of Rogers and Swaminathan (1993) as well as Uttaro and Millsap (1994), who found that the greatest impact on ES was for Type C items (i.e. items with large DIF).This notwithstanding, the finding in the present study that LR works best for Type C items compared to either Type B or Type C items does not concur with that of Zwick andErcikan (1989).

In a similar token, detection of DIF using LR statistic tends to improve slightly with Test Length, and this becomes more prominent with Type C items when ability distribution is Mean=1 SD=2. Indeed, differences in detection of DIF across item Types was more manifest in longer tests than shorter ones, with Type C items generally associated with the highest detection rates.

Conclusion

The effects of Sample Size, Ability Distribution and Test Length on ES of DIF items using Logistic Regression statistic was studied. Item responses were simulated for focal and reference groups, where the two groups had different ability distributions. The finding that Sample Size had a statistically significant effect on the ES for Type A items and not Type B or C items, and that Ability Distribution also had a statistically significant effect on the ES of Type C items and not for Type A or B items is a clear indication of the importance of making selective use of LR statistic in detecting DIF.

The finding that detection of DIF using LR statistic generally improves with Test Length regardless of the nature of Ability Distribution and Sample Size considerations confirms that longer tests are normally more desirable than shorter ones. This notwithstanding, such detection when LR statistic is used is better achieved for Type C items than either Type A or B items.

Recommendations

The following are recommendations based on the findings of the study:

(i) Test developers should pay more attention to Sample Size when measuring ES of DIF using LR procedure. This is more particularly so for A Type items (i.e. Items with Negligible DIF).

(ii) Test developers should consider Ability Distribution when using LR statistic to detect DIF. This is more particularly so for Type C items (i.e. Items with Large DIF).

Suggestions for Further Research

The following are suggestions for further research:

(i) Research on LR statistic focusing on polytomously scored items.

(ii) Research on the accuracy of LR statistic involving the independent variables used in the present study but with different levels.

(iii) Research exploring the accuracy of other methods of detecting DIF (e.g. SIBTEST) using the same independent variables.

(iv) Research comparing the accuracy of LR statistic and other DIF detection methods.

References

Cromwell, S.D. (2006). *Improving the Prediction of Differential Item Functioning: A comparison of the use of an Effect size for Logistic Regression DIF and Mantel-Haenszel DIF methods*. (Doctoral Dissertation), Texas A&M University.

Davies, J. P., Eisenhardt, K. M. & Bingham, C. B. (2007). Developing theory through simulation methods. *Academy of Management Review, 32*(2), 480-499.

Fidalgo, Á. M., Ferreres, D. & Muñiz, J. (2004**).** Liberal and conservative Differential Item Functioning detection using Mantel-Haenszel and SIBTEST: Implications for Type I and Type II error rates. *Journal of Experimental Education, 73*(1), 23-39. Retrieved on 17th January, 2008 from http://www.mendeley.com/.../angel-m-fidalgo/

González-Romá, V., Hernández, A., & Gómez-Benito, J. (2006). Power and Type I error of the mean and covariance structure analysis model for detecting differential item functioning in graded response items**.** *Multivariate Behavioral Research*, 41(1), 29- 53.

Hidalgo, M. D., & Lopez-Pina, J.A. (2004). Differential item functioning detection and effect size: a comparison between Logistic Regression and Mantel-Haenszel 137 procedures. *Educational and Psychological Measurement*, 64(6), 903-915. DOI: 10.1177/0013164403261769.

Han, K. T., & Hambleton, R. K. (2009). User's manual for WinGen: Windows software that generates IRT model parameters and item responses. Center for Educational Assessment Research Report No. 642. University of Massachusetts.

Harrison, J. R., Zhiang, L. I. N., Carrol, G. R. &Carley, K. M. (2007). Simulation modeling in organizational and management research. *Academy of Management Review, 32*(4), 1229-1245.

Holland, P.W., & Thayer, H. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp.129-145). Hillsdale, NJ: Lawrence Erlbaum Associates. Retrieved in2009fromhttp://www.books.google.co.ke/books?isbn*=1109 103204*

Jodoin, M. G., &Gierl, M.J. (2002). Evaluating type I error and power rates using an effect size measure with the Logistic Regression procedure for DIF detection. *Applied Measurement in Education*, 14, 329-349. Retrieved on 4th of November 2011 fromhttp://www.tandfonline.com/doi/full/*10.1080/15305058.2 011.60281*

Kathleen, M. M., Clauser, B. E. & Hambleton, R.K. (1992). The Effect of Sample Size on the Functioning of the Mantel-Haenszel Statistic. *Educational and Psychological Measurement*, *52*(2), 443-451.Retrieved on 30th March 2017 fromhttp://journals.sagepub.com/doi/abs/10.1177/0013164492 052002020

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. Retrieved on17th April, 2013 from www.*prezi.com/m1u58qcnpxbc/untitled-prezi/*

McCarthy, F. A., Oshima, T. C., & Raju, N.S. (2007). Identifying Possible Sources of Differential Functioning Using Differential Bundle Functioning with Polytomously Scored Data**.** *Applied Measurement in Education*, 20(2), 205–225 Retrievedin2011fromhttp**:**//education.gsu.edu/coshima/.../McCarty

Othuon, L. O. A. (1998). *The accuracy of parameter estimates and coverage probability of population values in regression models upon different treatments of systematically missing data.* Unpublished PhD thesis. University of British Columbia.

Pedrajita, Q J., & Talisayon, V.M. (2009). Identifying Biased Test Items by Differential Item Functioning Analysis Using Contingency Table Approaches: A Comparative Study. Education Quarterly, *University of the Philippines College of Education*, Vol. 67 (1), 21-43.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of Logistic Regression and Mantel-Haenszel procedures for detecting differential item functioning**.** *Applied Psychological Measurement*, 17, 105-116. Retrieved on 21st April 2013 from http://apm.sagepub.com/content/17/2/105.refs

Swaminathan, H., & Rogers, H. J. (1990**).** Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27, 361-370. Retrieved on 2nd March 2011 from *http://*www.jstatsoft.org/v39/i08/*paper*

Uttaro, T. & Millsap, R. E. (1994). Factors influencing the Mantel-Haenszel procedure in the detection of differential item functioning. *Applied Psychological Measurement*, *18,* 15-25.

Wang, W., & Su, Y. (2004). Factors influencing the Mantel and generalized Mantel-Haenszel methods for the assessment of Differential Item Functioning in polytomous items. *Applied Psychological Measurement*, *28*(6), 450-480. Retrieved on 4th May2012fromhttp://www.apm.sagepub.com/content/34/3/166.*refs*

Zieky, M. (1993). *Practical questions in the use of DIF statistics in test development*. In P. Holland & H. Wainer (Eds.), Differential item functioning (pp. 337–348). Hillsdale, NJ:Erlbaumhttp://scholar.lib.vt.edu/theses/.../etd.../RAA_ETD .pdf

Zumbo, B.D., & Thomas, D.R. (1996). *A measure of DIF effect size using logistic regression procedures***.** Paper presented at the National Board of Medical Examiners, Philadelphia. Retrieved on 19th September 2012 from http://www.educ.ubc.ca/faculty/**zumbo**/cv.htm

Zumbo, B. D. (1999). Logistic Regression Modeling as a unitary framework for Binary and Likert-type (ordinal) Item scores**.** *A Handbook on the Theory and Methods of Differential Item Functioning* (DIF) Ottawa, Canada, K1A 0K2.Retrieved19thSeptember,2012fromhttp://www.educ.ubc.ca/faculty/**zumbo**/cv.htm

Zwick, R. & Ercikan, K. (1989). Analysis of Differential Item Functioning in the NAEP History Assessment. *Journal of Educational Measurement, 26*(1), 55-66. Retrieved on 19thMarch2017fromhttp://carmeeduc.sites.olt.ubc.ca/files/2015/11/Zwick-Ercikan-1989.pdf