EFFECT OF SAMPLE SIZE, ABILITY DISTRIBUTION AND TEST LENGTH ON DETECTION OF DIFFERENTIAL ITEM FUNCTIONING USING MANTEL-HAENSZEL 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

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 Mantel-Haenszel (MH) 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 MH 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 (60*r*/60*f*) and a large sample size (1000*r*/1000*f*) 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 MH was used, test length had no statistically significant effect on the ES of DIF items. However, the number of DIF detections using MH 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.

Key words*: Differential Item Functioning (DIF), Mantel Haenszel (MH), effect size (ES), sample size, ability distribution, test length, WinGen3.*

INTRODUCTION

Background to the Study

Differential Item Functioning (DIF) is defined as the different probability of giving the right answer to a test item by two individuals with the same ability level, but from different groups (MaCarthy, Oshima & Raju, 2007). DIF can be determined by comparing two subpopulations' outcome on an item and also involves a decision of whether there is a large enough difference between subpopulations to eliminate or change the item of interest. 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).

* Corresponding author.

The Mantel-Haenszel (MH) procedure has become one of the most popular procedures for detecting differential item functioning (Kathleen, Clauser, & Hambleton, 1992). Rogers and Swaminathan (1993) showed that test length had no significant influence on the power of the MH 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 is varied.

In their simulation study, the MH procedure missed 25 to 30% of the differentially functioning items when sample size of 2000 was used in each of the focal and reference group. When sample size was reduced to 500 or fewer in each of the focal and reference group, more than 50% of the differentially functioning items were missed. The items most likely to be undetected were those which were most difficult, those with a small difference in item difficulty between the two groups, and poorly discriminating items.

The Mantel-Haenszel (MH) 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 is a non-parametric approach for identifying DIF (Mantel & Haenszel, 1959). MH is computed by matching examinees in each group on total test score and then forming a 2 (group) × 2 (item response) × K (score level), contingency table for each item where K is the score level on the matching variable of the total test score. At each score level j, a 2×2 contingency table is created for each item. The MH statistical procedure consists of comparing the item performance of two groups (reference and focal), whose members were previously matched on the ability scale. The matching is done using the observed total test score as a criterion or matching variable (Holland & Thayer, 1988). For dichotomous items, K contingency tables (2 ×2) are constructed for each item, where K is the number of test score levels into which the matching variable has been divided.

Under the MH procedure an effect size estimate based on the common odds ratio α is expressed as

$$\alpha_{MH} = \frac{\sum_{j=1}^{K} A_j \; D_{j/} \; / N_{..j}}{\sum_{j=1}^{K} B_j \; C_{j/} \; / N_{..j}}$$

Table 1 shows a 2 ×2 table for calculating the MH statistic for item *i* on a *j* score level in a test.

Table 1: Calculation of MH statistic for item *i* on a *j* score level in a test

Group	**1**	**0**	**Total**
Reference	A_j	B_j	$N_{R.j}$
Focal	C_j	D_j	$N_{F.j}$
Total	$\mathbf{N_{1j}}$	$\mathbf{N_{0j}}$	$\mathbf{N_{.j}}$

Holland and Thayer (1988) proposed a logarithmic transformation of α expressed as

$$\Delta\alpha_{MH} = -2.35\ln(\alpha_{MH})$$

Based on this transformation, Zwick and Ercikan (1989) proposed the following interpretation guidelines to evaluate the DIF effect size:

(i) Type A items – negligible DIF: items with $\Delta\alpha_{MH} < .|1|$

(ii) Type B items – moderate DIF: items with $\Delta\alpha_{MH} \geq |1|$ and $\leq |1.5|$ and the MH test is statistically significant.

(iii) Type C items – large DIF: items with $\Delta\alpha_{MH} > |1.5|$ and the MH test is statistically significant.

DIF is considered negligible if the magnitude $|\Delta - MH| < 1.5$. DIF is considered moderate when $\Delta - MH$ has either (a) $1 \leq |\Delta - MH| < 1.5$ or (b) $|\Delta - MH|$ is at least 1 but not significantly greater than 1. DIF is considered large when $\Delta - MH$ is significantly greater than 1 and $|\Delta - MH| \geq 1.5$ (Zieky, 1993). These ratings are referred to as A, B and C Types of DIF to denote negligible, moderate and large amounts of DIF, respectively. The purpose of this study was to investigate the effect of sample size, ability distribution and test length on detection of differential item functioning (DIF) using Mantel-Haenszel statistic.

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.

(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.

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. A routine was written, according to the MH formulae, which gave the effect size for MH analysis. The procedure was replicated 1000 times and the average effect size value was determined.

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 is presented in Table 2. 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 2: 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	.2355	1.2605	2.9469
10	(1, 2)	20	.6862	1.2053	4.8528
10	(0, 1)	60	.7964	1.0250	4.4606
10	(1, 2)	60	.4636	1.0596	5.5727
10	(0, 1)	1000	.1644	1.2986	2.3936
10	(1, 2)	1000	.5530	1.3772	3.4000
30	(0, 1)	20	.4857	1.2485	3.7856
30	(1, 2)	20	.8626	1.2322	4.2349
30	(0, 1)	60	.7735	1.2953	2.9986
30	(1, 2)	60	.6616	1.1273	4.7330
30	(0, 1)	1000	.5664	1.2431	3.3960
30	(1, 2)	1000	.6434	1.3500	7.3604
50	(0, 1)	20	.5655	1.2815	3.2351
50	(1, 2)	20	.8907	1.2000	5.1542
50	(0, 1)	60	.7935	1.2595	2.7136
50	(1, 2)	60	.6003	1.2601	4.0831
50	(0, 1)	1000	.5544	1.2356	3.7119
50	(1, 2)	1000	.4573	1.2934	4.7178

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 MH 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 3 summarizes the ANOVA results for the effect of Sample Size on the ES of DIF across 3 DIF Types using MH statistic.

Table 3: ANOVA Summary for effect of sample size on effect size of DIF across 3 DIF types

Type of DIF		Sum of Squares	df	Mean Square	F	Sig.
A	Between Groups	.115	2	.058	1.605	.234
	Within Groups	.539	15	.036		
	Total	.654	17			
B	Between Groups	.050	2	.025	4.234	.035
	Within Groups	.088	15	.006		
	Total	.137	17			
C	Between Groups	.050	2	.025	.015	.985
	Within Groups	24.718	15	1.648		
	Total	24.767	17			

Statistically significant differences between means was recorded for the B Type of DIF only (F=4.234, dfw=2, dfb=15, p=.035). Post-hoc analysis using Bonferroni method for pairwise comparisons revealed that for B Type DIF items, differences existed between sample size 60 and 1000 only as displayed in Table 4.

Table 4: Pairwise comparisons of effect sizes across different test lengths for Type B DIF
Dependent Variable: Effect Size
Post-hoc test: Bonferroni

(I) Sample size	(J) Sample size	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
20	60	.0668720	.0441727	.453	-.052118	.185862
	1000	-.0616358	.0441727	.550	-.180626	.057354
60	20	-.0668720	.0441727	.453	-.185862	.052118
	1000	-.1285079*	.0441727	.032	-.247498	-.009518
1000	20	.0616358	.0441727	.550	-.057354	.180626
	60	.1285079*	.0441727	.032	.009518	.247498

* The mean difference is significant at the 0.05 level.

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.

Table 4 summarizes the ANOVA results for the effect of Ability Distribution on the ES of DIF across 3 DIF Types using MH statistic.

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

Type of DIF		Sum of Squares	df	Mean Square	F	Sig.
	Between Groups	.043	1	.043	1.136	.302
A	Within Groups	.610	16	.038		
	Total	**.654**	**17**			
	Between Groups	.000	1	.000	.012	.915
B	Within Groups	.137	16	.009		
	Total	**.137**	**17**			
	Between Groups	11.627	1	11.627	14.158	.002
C	Within Groups	13.140	16	.821		
	Total	**24.767**	**17**			

Statistically significant differences for the effect of Ability Distribution on ES was recorded for C Type of DIF only (F_{obs}.=14.158, df_w=1, df_b=16, p=.002).

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 MH 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 5: ANOVA Summary for effect of test length on effect size of DIF across 3 DIF types

Type of DIF		Sum of Squares	df	Mean Square	F	Sig.
	Between Groups	.119	2	.059	1.668	.222
A	Within Groups	.535	15	.036		
	Total	**.654**	**17**			
	Between Groups	.009	2	.005	.541	.593
B	Within Groups	.128	15	.009		
	Total	**.137**	**17**			
	Between Groups	.926	2	.463	.291	.751
C	Within Groups	23.841	15	1.589		
	Total	**24.767**	**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.

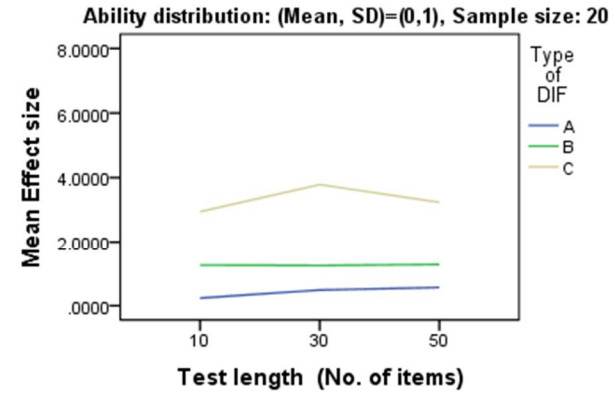
The largest mean ES was recorded for Type C DIF items. This was followed by Type B and C, respectively. This outcome was regardless of Ability Distribution, Sample Size and Test Length. However, differences in ES between B and C items were not as large as those between either A and B or A and C items.

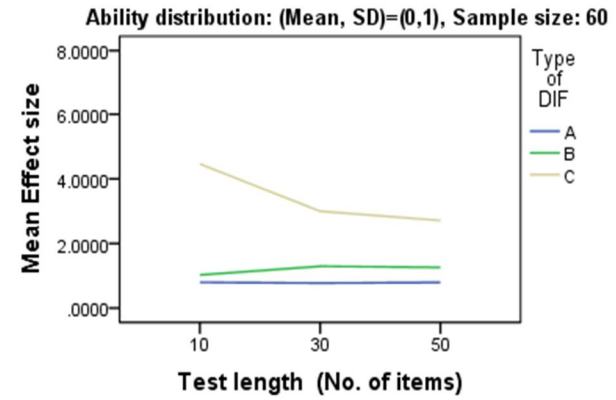
More specifically, for Ability Distribution with (Mean, SD)=(0, 1) and Sample Size=20, mean ES was largest for Type C items followed by B and A. However, the highest ES for Type C items occurred for 30 items. For Type C items, when Ability Distribution had (Mean, SD)=(1, 2) and Sample Size=20, the smallest 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 50 items with the magnitude of ES decreasing with Test Length. For Type A and B, ES tended to marginally increase with Test Length.

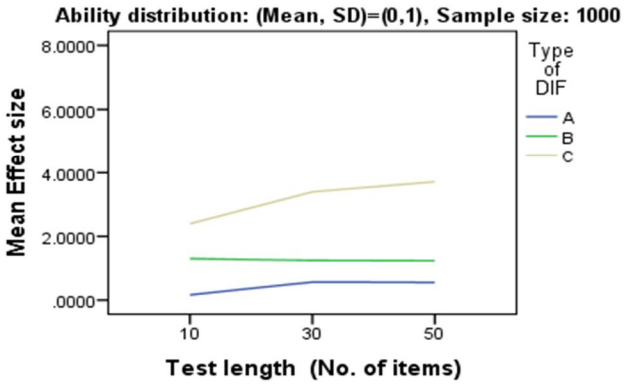
For Ability Distribution with (Mean, SD)=(0, 1) and Sample Size=60, the mean ES was largest for Type C items followed by B and A. For Type C DIF items in this category, the largest ES was recorded for 10 items and the smallest for 50 items with the magnitude of ES decreasing with Test Length. For Type A and B, ES tended to marginally increase with Test Length. This trend was reasonably maintained when the Ability Distribution with (Mean, SD)=(1, 2) 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 50 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.

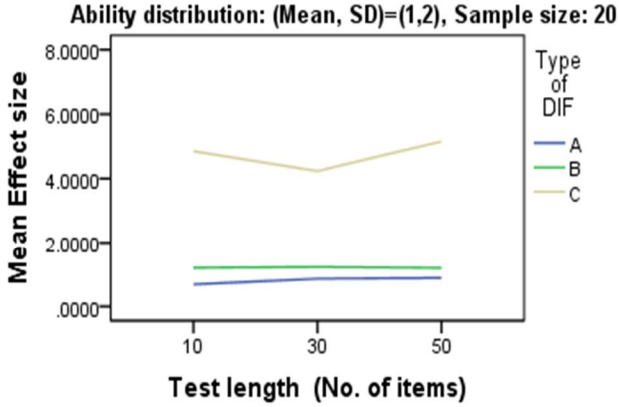
. **ABILITY DISTRIBUTION WITH MEAN=0, SD=1**

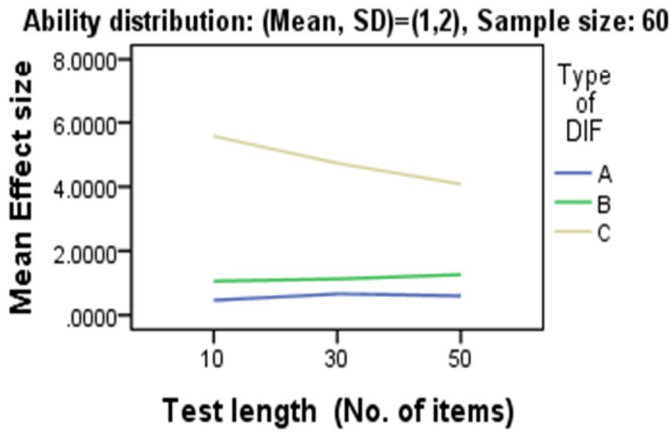






ABILITY DISTRIBUTION WITH MEAN=1, SD=2





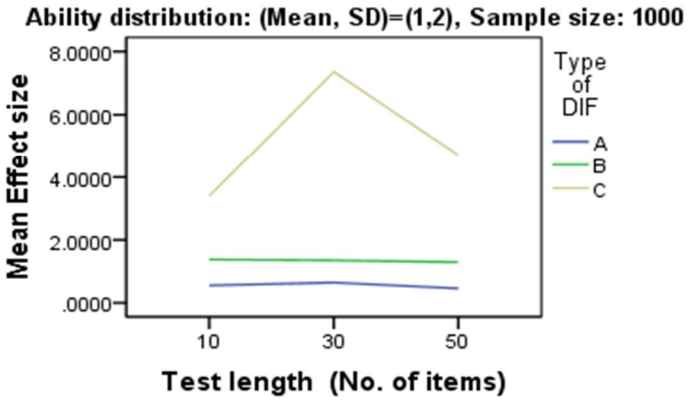


Figure 1: Mean effect sizes for different types of DIF under different conditions

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	0	4	6
10	(1, 2)	20	1	0	9
10	(0, 1)	60	1	1	8
10	(1, 2)	60	0	1	9
10	(0, 1)	1000	3	4	3
10	(1, 2)	1000	3	2	6
30	(0, 1)	20	0	8	22
30	(1, 2)	20	1	3	26
30	(0, 1)	60	5	4	21
30	(1, 2)	60	5	4	21
30	(0, 1)	1000	10	7	13
30	(1, 2)	1000	2	2	26
50	(0, 1)	20	0	23	27
50	(1, 2)	20	3	6	42
50	(0, 1)	60	16	13	21
50	(1, 2)	60	5	5	40
50	(0, 1)	1000	23	11	16
50	(1, 2)	1000	11	6	33

Key:

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

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

ABILITY DISTRIBUTION WITH MEAN=0, SD=1

ABILITY DISTRIBUTION WITH MEAN=1, SD=2

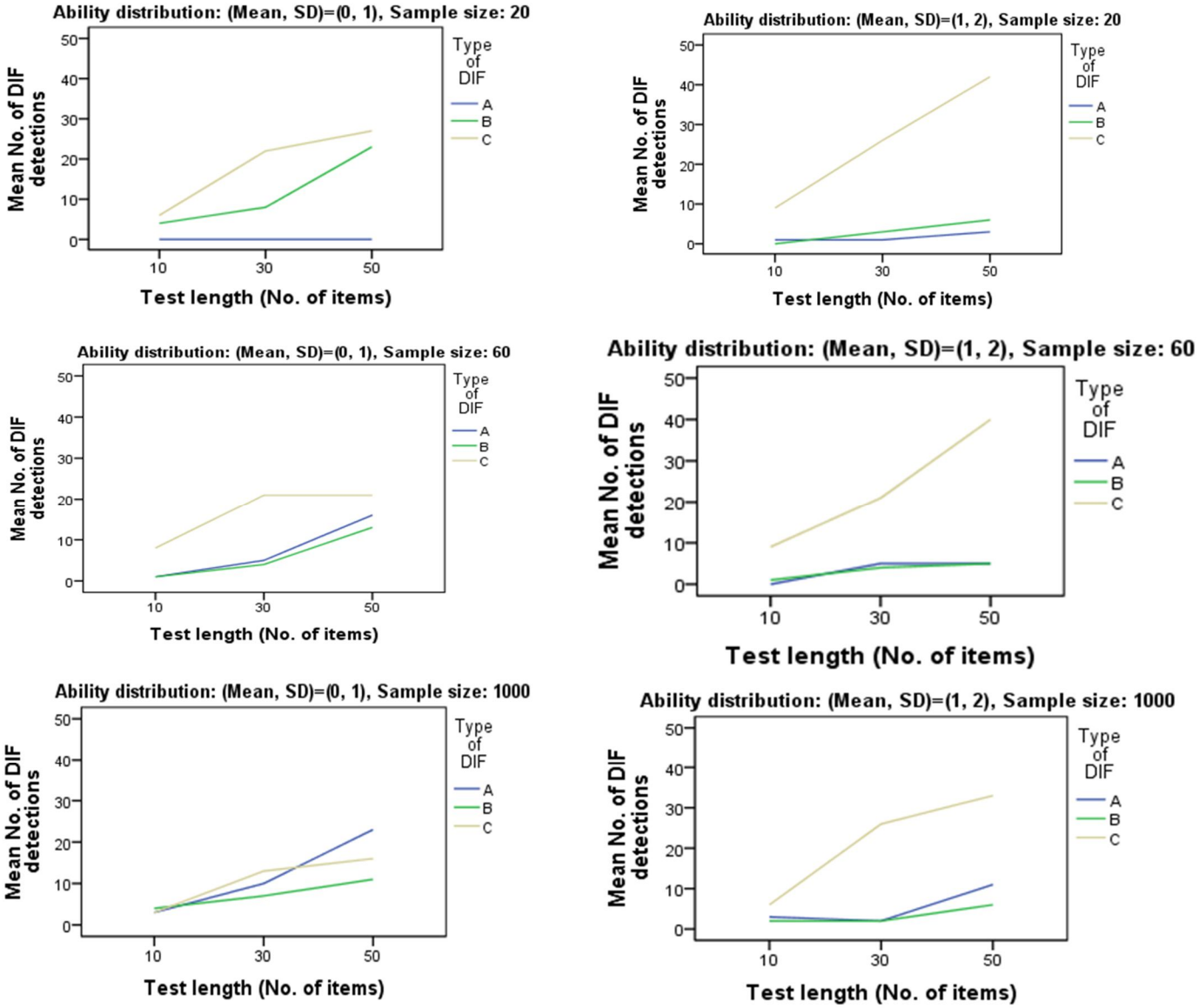


Figure 2: Mean number of DIF detections for different types of DIF under different conditions

In addition, the largest difference in DIF detection was recorded when Test Length was 30 items (Moderate 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 is (Mean, SD)=(0, 1), differences in mean DIF detection were minimal across the three types of DIF items i.e. A, B and C. However, differences in mean 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. A similar outcome was recorded for a sample of size 60, except that the difference in mean DIF detection for Type A and Type B items was minimal. 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 (50), MH 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 Mantel-Haenszel (MH) statistic. Results indicate that Sample Size had a statistically significant effect on ES for B Type items (Moderate DIF items) and not for A or C Types. Post-hoc test indicated that significant differences in ES for B Type items existed between Sample Size=60 and Sample Size=1000 only. This suggests that it is B Type items that may be problematic when measuring DIF using MH statistic, particularly for moderate 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 MH statistic.

Whereas Test Length had no statistically significant effect on ES for all the three item Types, there was a general trend for ES to increase with Test Length. This is consistent 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 MH works best for Type C items compared to either Type B or Type C items concurs with that of Zwick and Ercikan (1989).

In a similar token, detection of DIF using MH statistic tends to improve slightly with Test Length, and this becomes more prominent with Type C items. 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, **A**bility Distribution and Test Length on ES of DIF items using Mantel-Haenszel 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 B items and not Type A 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 MH statistic in detecting DIF.

The finding that detection of DIF using MH 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 MH 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 MH procedure. This is more particularly so for B Type items (i.e. Items with Moderate DIF).

(ii) Test developers should consider Ability Distribution when using MH 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 MH statistic focusing on polytomously scored items.

(ii) Research on the accuracy of MH 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. Logistic Regression) using the same independent variables.

(iv) Research comparing the accuracy of MH 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. (20017). 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-Hanszel 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 in 2009 from http://www.books.google.co.ke/books?isbn=*1109103204*

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 from http://www.tandfonline.com/doi/full/ *10.1080/15305058.2011.60281*

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 from http://journals.sagepub.com/doi/abs/10.1177/0013164492052002020

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 on 17th 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 Retrieved in 2011 from http***//***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.

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 May 2012 from http://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: Erlbaum

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 19th March 2017 from http://carme-educ.sites.olt.ubc.ca/files/2015/11/Zwick-Ercikan-1989.pdf