American Journal of Educational Research, 2019, Vol. 7, No. 11, 878-887
Available online at http://pubs.sciepub.com/education/7/11/19
Published by Science and Education Publishing
DOI:10.12691/education-7-11-19

Effectiveness of Mantel-Haenszel And Logistic Regression Statistics in Detecting Differential Item Functioning Under Different Conditions of Sample Size, Ability Distribution and Test Length

Ferdinand Ukanda[1], Lucas Othuon[1,*], John Agak[1], 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
*Corresponding author: lothuonus2013@gmail.com

Received October 10, 2019; Revised November 18, 2019; Accepted November 27, 2019

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) and Logistic Regression (LR) statistics provide effect size measures that quantify 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 number of DIF detections using MH and LR methods. 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. Small sample sizes (20*r*/20*f*), (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. The average effect sizes were obtained for 1000 replications. The number of DIF items were used to draw statistical graphs. The findings of the study showed that MH statistic detected more type A and B DIF items than LR regardless of the nature of Ability Distribution, Sample size and Test length. However MH statistic detected more type C DIF items than LR regardless of Ability Distribution, Sample size and Test length. The number of type C DIF items detected depended on the sample size, test length and ability distribution. Selective use of LR was therefore necessary for detecting type A and B DIF items while MH for detecting Type C DIF items. The findings of the study are of great significance to teachers, educational policy makers, test developers and test users.

Keywords: *Differential Item Functioning (DIF), Mante-Haenszel (MH), Logistic Regression (LR), effect size (ES), sample size, ability distribution, test length, WinGen3*

Cite This Article: Ferdinand Ukanda, Lucas Othuon, John Agak, and Paul Oleche, "Effectiveness of Mantel-Haenszel And Logistic Regression Statistics in Detecting Differential Item Functioning Under Different Conditions of Sample Size, Ability Distribution and Test Length." *American Journal of Educational Research*, vol. 7, no. 11 (2019): 878-887. doi: 10.12691/education-7-11-19.

1. Introduction

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 [1]. 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 (e.g., male vs. female Caucasian vs. African-American [2].

DIF can be determined by comparing two subpopulations' outcome on an item and also involve 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. DIF statistics that can provide an Effect size measure to be used to quantify the magnitude of DIF when detected include; the Mantel-Haenszel (MH) and Logistic Regression (LR) statistics.

The Mantel-Haenszel (MH) method has been one of the common methods for detecting differential item functioning [3]. 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 [4]. 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 [5]. 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

$$\alpha_{MH} = \frac{\sum_{j=1}^{K} A_j D_{j/} / N_{..j}}{\sum_{j=1}^{K} A_j C_{j/} / N_{..j}}. \quad (1)$$

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	N_{1j}	N_{0j}	$N_{.j}$

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

$$\Delta\alpha_{MH} = -2.35ln\,(\alpha_{MH}) \quad (2)$$

Based on this transformation, Zwick and Erickan [6] proposed the following interpretation guidelines to evaluate the DIF effect size: Type A items – negligible DIF: items with $|\Delta\alpha_{MH}| < .1$

Type B items – moderate DIF: items $|\Delta\alpha_{MH}| > 1$ and ≤ 1.5 and the MH test is statistically significant. Type C items – large DIF: items with $|\Delta\alpha_{MH}| >$dna 1.5 the MH test is statistically significant.

DIF was considered negligible if the magnitude $|\Delta\text{-}MH| < 1.5$. DIF was considered moderate when $\Delta\text{-}MH$ has either (a) $1 \leq |\Delta\text{-}MH| < 1.5$ or (b) $|\Delta\text{-}MH|$ is at least 1 but not significantly greater than 1. DIF is considered large when $\Delta\text{-}MH$ is significantly greater than 1 and $|\Delta\text{-}MH| \geq 1.5$ [7]. These ratings are referred to as A, B and C Types of DIF to denote negligible, moderate and large amounts of DIF, respectively.

The Logistic Regression (LR) method has also been used widely in DIF research [3]. The method also provides an effect size measure that quantifies the magnitude of DIF when detected. 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 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:

$$p(u=1) = \frac{e^z}{1+e^z} \quad (3)$$

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, $z = \beta_0 + \beta_1 X$ 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. $z = \beta_0 + \beta_1 X + \beta_2 G$) 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. $z = \beta_0 + \beta_1 X + \beta_2 G + \beta_3 XG$) subtracted from Model 2. Zumbo and Thomas [8] *d*eveloped an index to quantify the magnitude of DIF for the LR procedure based on partitioning a weighted least-squares estimate of R^2 that yields an effect size measure. This index is obtained, first, by computing the R^2 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 R^2 for each of the terms. A DIF effect size for the group membership term is produced by subtracting the R^2 for the group membership term (Model 2) from the R^2 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 $R^2\Delta$ - U). A second DIF effect size is produced for the total score-by-group membership term by subtracting the R^2 for` the group membership interaction that quantifies the magnitude of non-uniform DIF (herein called $R^2\Delta$ - N). $R^2\Delta$ can be used with the LR significance test to identify items with DIF. Jodoin [9] empirically-established guidelines for interpreting $R^2\Delta$. An item has negligible or A-level DIF when the chi-square test for model fit is not statistically significant or when $R^2\Delta < 0.035$. An item has moderate or B-level DIF when the chi square test is statistically significant and when $0.035 \leq R^2\Delta < 0.070$. An item has large or C-level DIF when the chi-square test is statistically significant and when $R^2\Delta \geq 0.070$. These guidelines are applicable to both uniform and non-uniform DIF, and were used to classify DIF items in the current study.

The aim of the comparison of the two methods was to determine if the effect of different Sample sizes, Test lengths and Ability distribution on the number of DIF detections of different types, was dependent on the procedure for DIF detection. For instance a simulation study by Salubayba [1] noted that different conditions such as Sample size affected the accuracy of some DIF detection methods. The study noted that below Sample size 100, DIF items were not detected using SIBTEST method. Hernandez and Gomez- Bento [10] used Sample

sizes of 100, 200, 400 and 800 which were perceived to be large enough to determine the effect of Sample size on DIF detection. Most of the studies used a Sample size more than 100 but less than 1000. These findings provided a basis for investigating whether a Sample size smaller than 100 say 60 or 20 and that larger than 100 say 1000, had any significant effect in the detection of DIF using two DIF statistics namely Mantel-Haenszel and Logistic Regression.

A study by Khalid [11] used items of varied Test lengths (40-80 items) and noted its effect on the accuracy of DIF detection. It was found that the influence of Test Length was rather modest and that the number of items did not greatly affect the detection of DIF of any kind. This finding provided a basis for investigating the effect of Test Length, using a test with as few items as 10 and as many items as 30 or 50. These numbers were considered basing on the number of items used in most real testing situations. Also to be considered is whether the effect of Test Length was dependent on the DIF detection procedure such as Mantel-Haenszel and Logistic Regression.

Studies have found that differences in Ability Distributions, assessed in terms of mean and standard deviation of the data, affected DIF detection rates [12]. They simulated data generated only from mean 0, standard deviation1 for both focal and reference groups. This provided a basis for comparing the effect in DIF detection using simulated data with mean 0, standard deviation1 and mean 1, standard deviation 2 using two DIF statistics namely Mantel-Haenszel and Logistic Regression.

1.1. Purpose of the Study

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 and Logistic Regression statistics.

1.2. 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 number of Type A DIF items using MH and LR statistics.

ii) Determine the effect of Sample Size, Ability Distribution and Test Length on the number of Type B DIF items using MH and LR statistics.

iii) Determine the effect of Sample Size, Ability Distribution and Test Length on the number of Type C DIF items using MH and LR statistics.

2. Methodology

2.1. 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 variable was the number of DIF items detected based on the magnitude of the effect sizes.

2.2. 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*).

2.3. Data Collection Procedure

WinGen3 [13] 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 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 to 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 data was obtained for 1,000 replications, 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.

2.4. Methods of Data Analysis

Analysis was done using the Statistical Package for Social Sciences (IBM SPSS Version 20) computer software. Analysis was done on the raw data in order to obtain the Effect sizes for both MH and LR methods. For the MH method analysis was done using a routine was written, according to the MH formulae on MS Excel computer operating system, which gave the Effect size for MH analysis. The procedure was repeated for 1000 replications and the average Effect size values were determined. The number of items displaying various categories of DIF were then determined.

For LR method, the Statistical Package for Social Sciences (SPSS) (IBM SPSS Version 20) was used for analysis using 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. Line graphs for mean number of items across various categories of DIF were constructed to aid interpretation.

3. Results

3.1. Number of Type A DIF Items under Different Conditions

The number of Type A DIF items detected under different conditions is presented in Table 2. The number of Type A DIF items were compared for MH and LR statistics under different conditions of Sample size, Ability distribution and Test length Line graphs showing the mean number of detections for Type A DIF under different conditions of Sample size, Ability distribution and Test length were compared for MH and LR statistics. Figure 1 shows the mean number of Type A DIF detections under different conditions using MH and LR statistics. From the graphs it can generally be seen that LR statistic detected more Type A DIF items than MH statistic regardless of the Sample size, Ability distribution and Test Length. This indicates that LR is a better statistic for detecting Type A DIF than MH. When the Ability Distribution was such that (Mean, SD)=(0, 1), and the Sample Size was 20, only small differences in DIF detection occurred for Type A items between MH and LR statistics for 10 items while large differences occurred for 50 items. The number of items detected remained the same regardless of the Test length for MH statistic while it increased with Test length for the LR statistic. When the Ability Distribution was such that (Mean, SD)=(0, 1), and the Sample Size was 60, the number of DIF detections increased with Test length regardless of the DIF statistic.

Table 2. Number of Type A DIF items detected under different conditions for MH and LR statistics

No. of items	Ability distribution (Mean, SD)	Sample size	Number of DIF detections	
			MH	LR
10	(0, 1)	20	0	3
10	(1, 2)	20	1	2
10	(0, 1)	60	1	2
10	(1, 2)	60	0	7
10	(0, 1)	1000	3	9
10	(1, 2)	1000	3	5
30	(0, 1)	20	0	11
30	(1, 2)	20	1	9
30	(0, 1)	60	5	13
30	(1, 2)	60	5	9
30	(0, 1)	1000	10	18
30	(1, 2)	1000	2	7
50	(0, 1)	20	0	18
50	(1, 2)	20	3	12
50	(0, 1)	60	16	34
50	(1, 2)	60	5	13
50	(0, 1)	1000	23	32
50	(1, 2)	1000	11	21

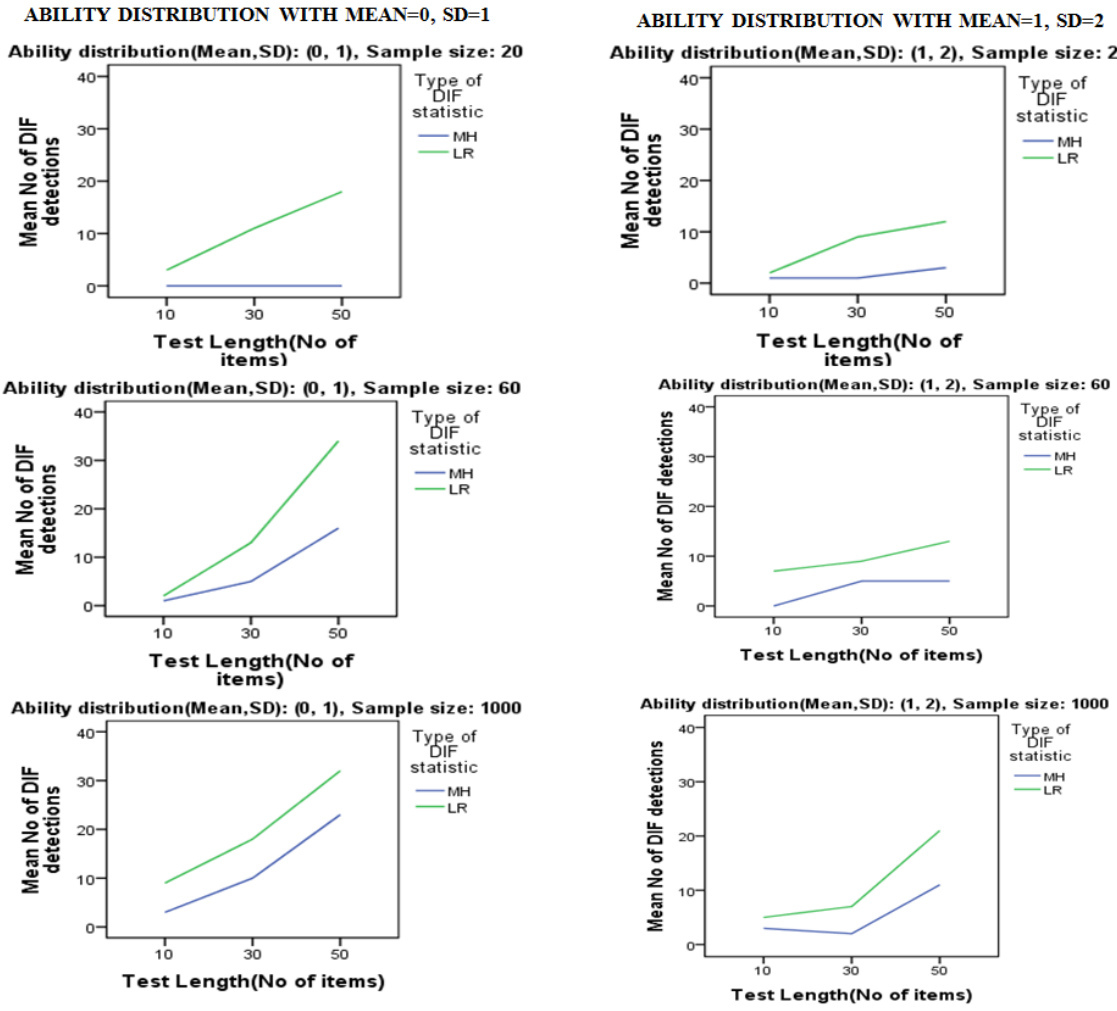

Figure 1. Mean number of DIF detections for Type A DIF under different conditions using MH and LR statistics

However LR detected more Type A items than MH with the highest number detected for a Test length of 50 items. From the graphs it can be seen that LR statistic detected more Type A DIF items than MH statistic regardless of the Sample size, Ability distribution and Test Length. The number of items detected by the MH statistic by was almost the same regardless of the Test length while it increased with Test length for the LR statistic. When the Ability Distribution was such that (Mean, SD)=(0, 1), and the Sample Size was 60, the number of DIF detections increased with Test length regardless of the DIF statistic. However LR detected more Type A items than MH with the highest number detected for a Test length of 50 items. When the Ability Distribution was such that (Mean, SD)=(0, 1), and the Sample Size was 1000, the number of DIF detections increased with Test length regardless of the DIF statistic. However LR detected more Type A items than MH with the highest number detected for a test length of 50 items. This number was however lower than that when the Sample size was 60.

When the Ability Distribution was such that (Mean, SD)=(1, 2), and the Sample Size was 20, the number of DIF detection increased with Test length regardless of the LR DIF statistic but remained at a low level for MH statistic. The number of DIF detections were however lower than when the Ability Distribution was such that (Mean, SD)=(0, 1). This indicated that Ability distribution had an effect on the number of DIF detections for LR but not for MH statistic. However LR detected more Type A items than MH with the highest number detected for a test length of 50 items. When the Ability Distribution was such that (Mean, SD)=(1, 2), and the Sample Size was 60, the number of DIF detections was lower than when Ability Distribution was such that (Mean, SD)=(0, 1), for 30 and 50 items for both MH and LR statistics. A large difference was noted in the number of DIF detections between MH and LR statistics for 30 and 50 items. This further indicated that Ability distribution and Test length had an effect on the number of DIF detections by both MH and LR statistics. However LR detected more Type A items than MH with the highest number detected for a Test length of 50 items. When the Ability Distribution was such that (Mean, SD)=(1, 2), and the Sample Size was 1000, the number of DIF detections decreased with Test length for 30 items using the MH statistic. As earlier noted the number of DIF detections were less than when the Ability distribution was such that (Mean, SD)=(0,1) for both statistics.

It was therefore noted that LR detected more Type A DIF items than MH statistic regardless of the Ability distribution, Test length and Sample size. However more detections were noted when the Ability distribution was such that (Mean, SD)=(0, 1) than when the Ability distribution was such that (Mean, SD)=(1, 2) for both MH and LR statistics. Also the number of DIF detections increased with Test length regardless of the Sample size and Ability distribution.

3.2. Number of Type B DIF Items under Different Conditions

The number of Type B DIF items detected under different conditions is presented in Table 3. The number of Type B DIF items were compared for MH and LR statistics under different conditions of Sample size, Ability distribution and Test length. Line graphs showing the mean number of detections for Type B DIF under different conditions of Sample size, Ability distribution and Test length were compared for MH and LR statistics. Figure 2 shows the mean number of Type B DIF detections under different conditions using MH and LR statistics.

From the graphs it can be seen that LR statistic detected more Type B DIF items than MH statistic regardless of the Sample size, Ability distribution and Test Length. When the Ability Distribution was such that (Mean, SD)=(0, 1), and the Sample Size was 20, only small differences in DIF detection occurred for Type B items between MH and LR statistics for 10 items while large differences occurred for 50 items.

Table 3. Number of Type B DIF items detected under different conditions for MH and LR statistics

No. of items	Ability distribution (Mean, SD)	Sample size	Number of DIF detections	
			MH	LR
10	(0, 1)	20	4	4
10	(1, 2)	20	0	0
10	(0, 1)	60	1	3
10	(1, 2)	60	1	0
10	(0, 1)	1000	4	3
10	(1, 2)	1000	2	8
30	(0, 1)	20	8	5
30	(1, 2)	20	3	5
30	(0, 1)	60	4	9
30	(1, 2)	60	4	4
30	(0, 1)	1000	7	1
30	(1, 2)	1000	2	17
50	(0, 1)	20	23	6
50	(1, 2)	20	6	8
50	(0, 1)	60	13	8
50	(1, 2)	60	5	8
50	(0, 1)	1000	11	4
50	(1, 2)	1000	6	5

The number of items detected was almost the same regardless of the Test length for MH statistic while it increased with Test length for the LR statistic. This result was similar to that of Type A DIF items for the same Ability distribution. When the Ability Distribution was such that (Mean, SD)=(0, 1), and the Sample Size was 60, the number of DIF detections increased with Test length regardless of the DIF statistic. However LR detected more Type B items than MH with the highest number detected for a Test length of 50 items. This result was also similar to that of Type A DIF items for the same Ability distribution. When the Ability Distribution was such that (Mean, SD)=(0,1), and the Sample Size was 1000, the number of DIF detections increased with Test length regardless of the DIF statistic. However LR detected more Type B items than MH with the highest number detected for a Test length of 50 items. This number was however lower than that when the Sample size was 60. The result showed more Type B items detected for Test length of 10 items than Type A DIF items.

When the Ability Distribution was such that (Mean, SD)=(1, 2), and the Sample Size was 20, the number of

Type B DIF detections increased with Test length for the LR DIF statistic but remained at a low level for MH statistic. The number of DIF detections were however lower than when the Ability Distribution was such that (Mean, SD)=(0, 1). This indicated that Ability distribution had an effect on the number of DIF detections for LR but not for MH statistic. However LR detected more Type B items than MH with the highest number detected for a Test length of 50 items. When the Ability Distribution was such that (Mean, SD)=(1, 2), and the Sample Size was 60, the number of DIF detections for LR increased with Test length while that for MH decreased from 30 to 50 items. This indicated that Test length had a significant effect on the detection of DIF for MH statistics. However LR still detected more DIF items regardless of the Test length with the greatest number being detected for a Test length of 50. This further indicated that Ability distribution and Test length had an effect on the number of DIF detections by both MH and LR statistics. When the Ability Distribution was such that (Mean, SD)=(1, 2), and the Sample Size was 1000, the number of DIF detections decreased with Test length from 10 to 30 items using the MH statistic and then increased with a Test length of 50 . For LR the mean number of DIF items detected increased with the Test length. As earlier noted the number of DIF detections for Sample size 1000 were less than when the Ability distribution was such that (Mean,SD)=(0,1) for both statistics. This further indicated that Ability distribution had an effect on the detection of Type B DIF items for both the MH and LR DIF statistics.

It was therefore noted that LR detected more Type B DIF items than MH statistic regardless of the Ability distribution, Test length and Sample size. However more detections were noted when the Ability distribution was such that (Mean, SD)=(0, 1) than when the Ability distribution was such that (Mean, SD)=(1, 2) for both MH and LR. Also the number of DIF detections increased with Test length regardless of the Sample size and Ability distribution. It was also noted that when the Ability distribution was such that (Mean, SD)=(1, 2) the number of DIF detections for MH statistic was less for Test length 30 than that of Test length 10 but increased when the Test length was 50.

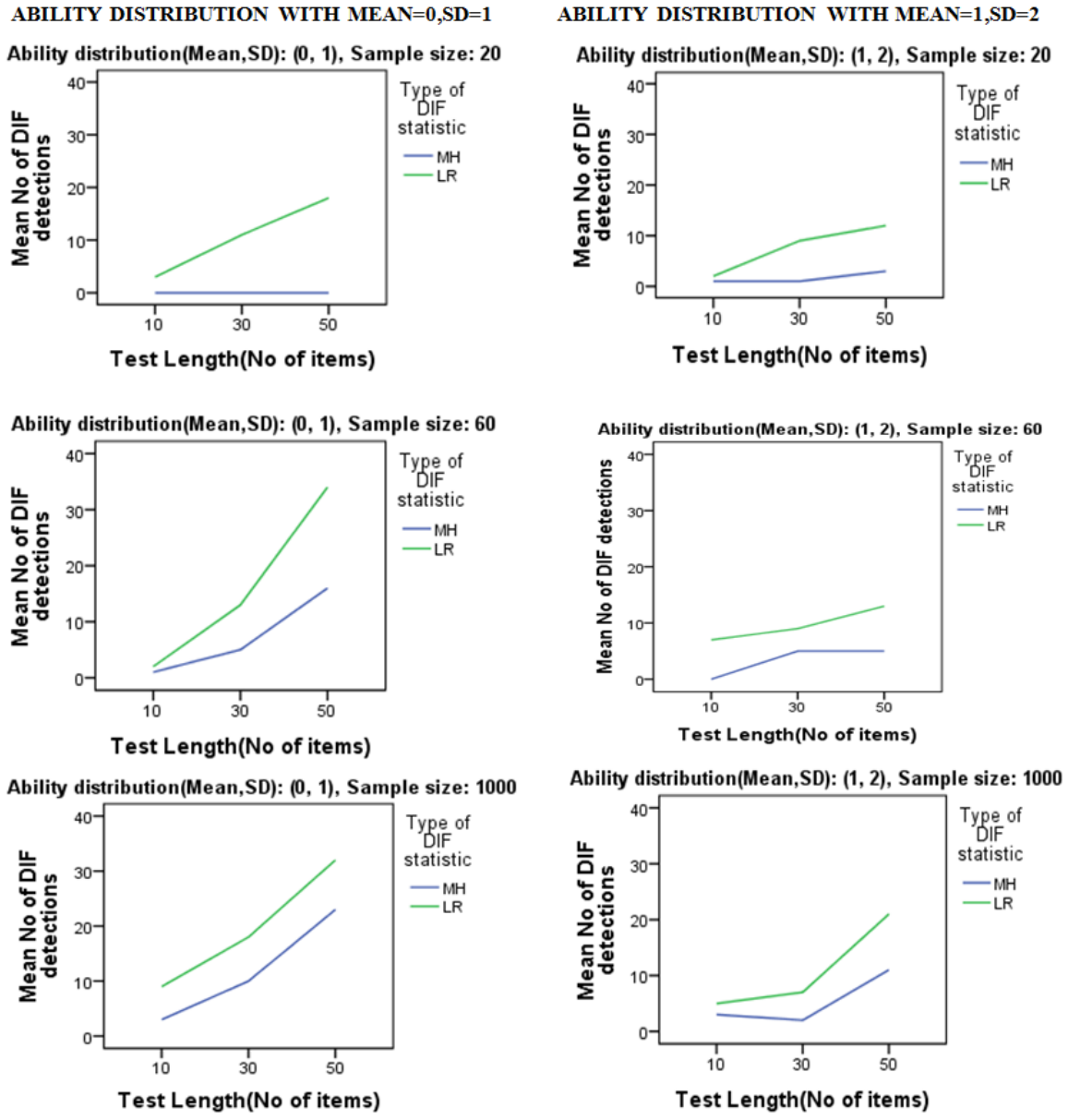

Figure 2. Mean number of DIF detections for Type B DIF under different conditions using MH and LR statistics

3.3. Number of Type C DIF Items under Different Conditions

The number of Type C DIF items detected under different conditions is presented in Table 4.

The numbers of Type C DIF items were compared for MH and LR statistics under different conditions of Sample size, Ability distribution and Test length. Line graphs showing the mean number of detections for Type C DIF under different conditions of Sample size, Ability distribution and Test length were compared for MH and LR statistics. Figure 3 shows the mean number of Type C DIF detections under different conditions using MH and LR statistics. Figure 3 shows the mean number of Type C DIF detections under different conditions using MH and LR statistics. From the graphs it can be seen that the MH statistic detected more Type C DIF items than LR statistic regardless of the Sample size, Ability distribution and Test Length. When the Ability Distribution was such that (Mean, SD)=(0, 1), and the Sample Size was 20, only small differences in DIF detection occurred for Type C items between MH and LR statistics for 10 items while large differences occurred for 30 and 50 items.

Table 4. Number of Type C DIF items detected under different conditions for MH and LR statistics

No. of items	Ability distribution (Mean, SD)	Sample size	Number of DIF detections	
			MH	LR
10	(0, 1)	20	6	3
10	(1, 2)	20	9	8
10	(0, 1)	60	8	5
10	(1, 2)	60	9	2
10	(0, 1)	1000	3	1
10	(1, 2)	1000	6	2
30	(0, 1)	20	22	11
30	(1, 2)	20	26	16
30	(0, 1)	60	21	12
30	(1, 2)	60	21	12
30	(0, 1)	1000	13	8
30	(1, 2)	1000	26	22
50	(0, 1)	20	27	15
50	(1, 2)	20	42	32
50	(0, 1)	60	21	8
50	(1, 2)	60	40	29
50	(0, 1)	1000	16	14
50	(1, 2)	1000	33	24

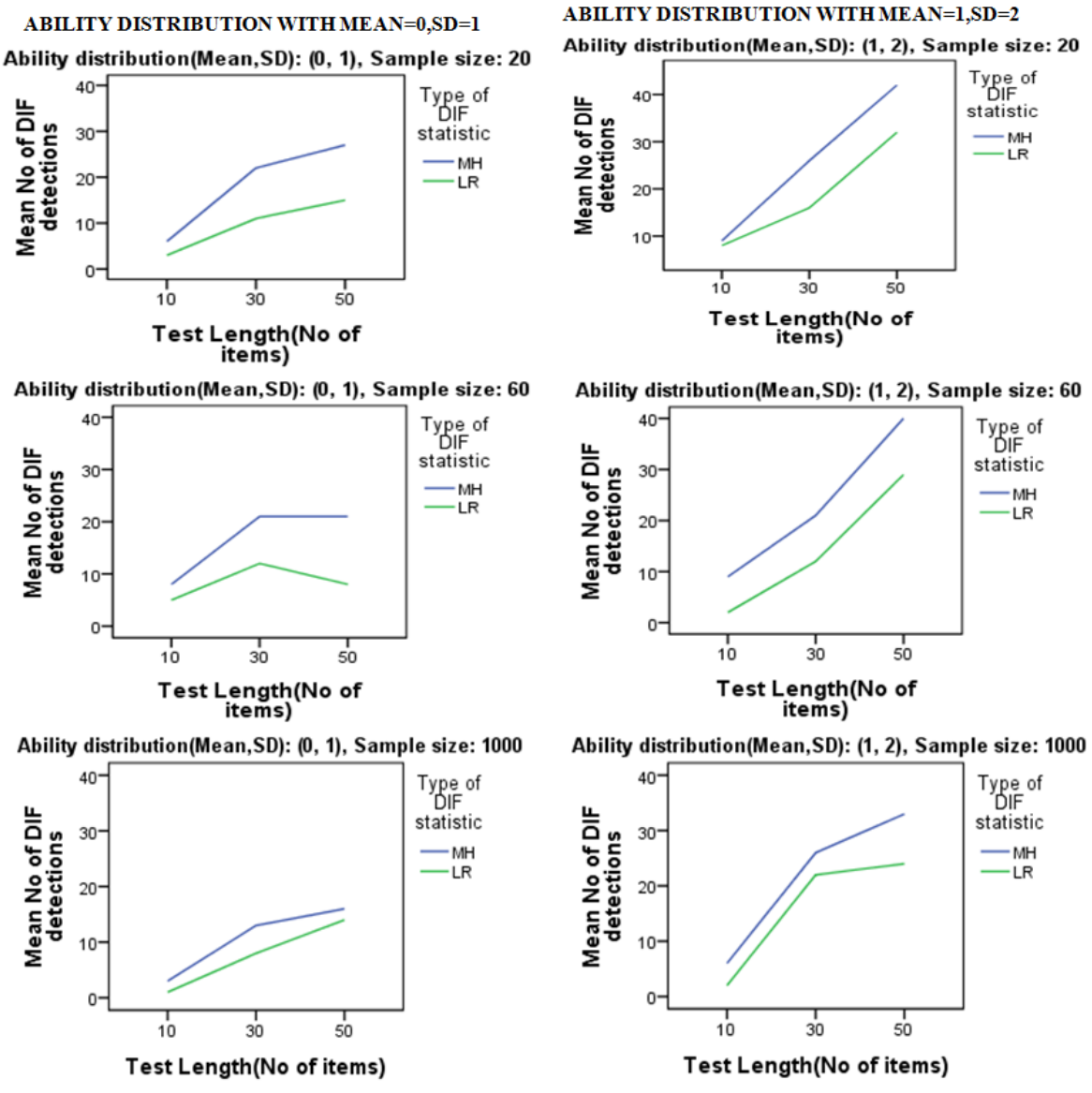

Figure 3. Mean number of DIF detections for Type C DIF under different conditions using MH and LR statistics

The number DIF items detected increased with Test length for both MH and LR statistics When the Ability Distribution was such that (Mean, SD)=(0, 1), and the Sample Size was 60, the number of Type C DIF detections increased with Test length between 10 to 30 items and was the same for 30 and 50 items using MH the statistic. For LR statistic, the number of Type C DIF items detected increased with Test length between 10 to 30 items and decreased with Test length between 30 and 50 items. When the Ability Distribution was such that (Mean, SD)=(0, 1), and the Sample Size was 1000, the number of DIF detections increased with Test length regardless of the DIF statistic ,small differences occurred between MH and LR in the number of Type C DIF items detected. When the Ability Distribution was such that (Mean, SD)=(1,2) Small differences in the Type C DIF detection between MH and LR statistics occurred for Test length 10 while large differences occurred for 30 and 50 items. When the Ability Distribution was such that (Mean, SD)=(1, 2), and the Sample Size was 60, the number of Type C DIF detections increased with Test length for both MH and LR statistics.

Large differences in the DIF detection occurred between the two statistics with the largest difference occurring for 50 items. When the Ability Distribution was such that (Mean, SD)=(1, 2), and the Sample Size was 1000, the number of Type C DIF detections increased with Test length for both statistics. The difference in the detection of Type C DIF items between MH and LR statistics was small for a Test length of 10 and 30 but it was large for a Test length of 50 items. This indicated that Test length had an effect on the number of Type C DIF detections using the MH and LR statistics.

It was therefore noted that MH detected more Type C DIF items than LR statistic regardless of the Ability distribution, Test length and Sample size. However more Type C DIF detections were noted when the Ability distribution was such that (Mean, SD)=(1, 2) than when the Ability distribution was such that (Mean, SD)=(0, 1) for both MH and LR statistics. Also the number of DIF detections increased with Test length regardless of the Sample size and Ability distribution in some instances while it decreased with Test length in other instances. Ability distribution, Test length and Sample size therefore had an effect on the number of Type C DIF detections by both the MH and LR statistics.

3.4. 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, Zhiang, Carrol and Carley [14] argue that simulation is prone to misspecification errors. Further, Davies, Eisenhardt and Bingham [15] 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 [16], and Davies, Eisenhardt and Bingham [15] 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.

4. Discussion

The purpose of this study was to investigate the effect of Sample Size, Ability Distribution and Test Length on number of DIF items detected using Mantel-Haenszel (MH) and Logistic Regression (LR) statistics.

The objectives were to compare the effect of Sample size, Ability Distribution and Test Length on the number of Type A, B and C DIF detections using MH and LR Statistics. The findings indicate that the Sample size had a small effect on the detection of Type A DIF items. Ability distribution did have an effect in the detection of Type A items, by both MH and LR statistics. The findings also indicate that the Sample size had a small effect on the detection of Type B DIF items, regardless of the Ability distribution using either MH or LR statistics. However Ability distribution did have an effect in the detection of Type B items, by both MH and LR statistics. Also that LR statistic detected more Type B DIF items than MH statistic notwithstanding the Ability distribution, Sample size and Test length. These findings are consistent with previous research by Hidaligo and Lopez Pina [17] which stated that Logistic regression analysis generally detected more items with DIF than the standard MH procedure. Their research did not indicate the category of DIF detected. They also stated that Test length (40 to 80 items), resulted in improved performance of the MH procedure which is consistent with the findings of this study. These findings are also consistent with previous research by Khalid [11] who examined the power of MH procedure by varying the magnitude of DIF, Test length and Sample size. It was found that the influence of Test length was rather low. The findings showed that the number of items do not greatly affect the detection of DIF of any kind by MH method. The results were not consistent with those of a study by Gonzalez-Romá, Hernandez and Gomez-Benito [18] who indicated that power of DIF statistics increased as Sample sizes and DIF magnitude increased and that the control for Type I error was better when sample sizes were large.

Test developers and test users can use the findings to make informed decisions regarding the selection of test item evaluation procedures in the area of differential item functioning under different examinee conditions.

The findings also indicate that the Sample size had a small effect on the detection of Type C DIF items of regardless of the Ability distribution using either MH or LR statistics. However Ability distribution did have an effect in the detection of Type C items, by both MH and LR statistics. Also that MH statistic detected more Type C DIF items than LR statistic notwithstanding the Ability distribution, Sample size and Test length. These findings are not consistent with previous research by Hidaligo and Lopez Pina [17] which stated that Logistic regression analysis generally detected more items with DIF than the standard MH procedure. Their research did not indicate the category of DIF detected. They also stated that Test

length (40 to 80 items), resulted in improved performance of the MH procedure. The findings were also not consistent with previous research by Pedrajita and Talisayon [19] who found a high degree of agreement between LR and MH statistics in identifying biased items(Type C DIF items). The study however used real data from Junior high school students from public and private schools. The findings were also not consistent with previous research by Adedoyin [20] which used IRT method to detect gender biased items in public examinations. The study found out that out of 16 test items that fitted the 3PL IRT analysis 5 were gender biased. The study used 2000 males and 2000 females which was a large Sample size. The findings were also not consistent with previous research by Fidalgo, Ferreres, & Muñiz, [21] who reported that DIF detection by either M-H or an IRT based procedure resulted in inflated Type I error.

The findings of this study can also contribute the formulation and implementation of educational policies and decisions related to test development.

5. Conclusion

The effects of Sample Size, Ability Distribution and Test Length on the number of DIF items detected using Mantel-Haenszel and Logistic Regression statistics were studied. Item responses were simulated for the focal and reference groups, where the two groups had different ability distributions. The finding that Logistic Regression statistic detected more Type A and B items and Mantel-Haenszel statistic detected more Type C items is a clear indication of the importance of making selective use of the LR statistic when detecting Type A and B items and the MH statistic when detecting Type C DIF items. Such detection was achieved regardless of the Sample size, Test Length and Ability distribution.

5.1. Recommendations

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

i) Test developers should pay more attention to using LR procedure particularly for detecting Type A and B items (i.e. Items with small and Moderate DIF).

ii) Test developers should consider using MH statistic to detect particularly Type C DIF items (i.e. Items with Large DIF).

5.2. Suggestions for Further Research

The following are suggestions for further research:

i) Research on MH and LR statistics focusing on polytomously scored items.

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

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

References

[1] Salubayba, T. M. Differential item functioning detection in reading comprehension test using Mantel-Haenszel, Item response Theory, and logical data analysis. The international Journal of social sciences, 2013, 14(1), 76-82.

[2] Schumacker, R. Test bias and differential item functioning,2005. Retrieved on 2nd March 2011 from http://www.appliedmeasurementassociates.com/White Papers/TEST Bias and Differential Item Functioning.pdf.

[3] Wang, W., & Su, Y. Factors influencing the Mantel and generalized Mantel-Haenszel methods for the assessment of Differential Item Functioning in polytomous items. Applied Psychological Measurement, 2004, 28(6), 450-480.

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

[5] Holland, P.W., & Thayer, H. Differential item performance and the Mantel-Haenszel procedure. In Weiner, H. & Braun, H. (Eds,). Test Validity. 1988, 129-145. Hillsdale, NJ: Laurence Erlbaum Associates. Retrieved in 2009 from http://www.books.google.co.ke/books?isbn=1109103204.

[6] Zwick, R.& Erickan, K. Analysis of Differential Item Functioning in the NAEP History Assessment. Journal of Educational Measurement, 1989, 26(1), 55-66.

[7] Zieky, M. Practical questions in the use of DIF statistics in test development. In P. Holland & H. Wainer (Eds.), Differential item functioning 1993, 337-348 . Hillsdale, NJ: Erlbaum.

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

[9] Jodoin, M. G., & Gierl, M.J. Evaluating type I error and power rates using an effect size measure with the Logistic Regression procedure for DIF detection. Applied Measurement in Education. 2002, 14, 329-349.

[10] Hernández, A., & González-Romá, V. Evaluating the multiple-group mean and covariance structure model for the detection of differential item functioning in polytomous ordered items. Psichtema, 2003, 15, 322-327.

[11] Khalid, N, M. The performance of Mantel-Haenszel procedures in the identification of DIF items. International Journal of Educational Sciences 2011, 3(2), 435-447.

[12] French, B. F., & Maller, S. J. Iterative purification and effect size use with Logistic Regression for differential item functioning. Educational Psychological Measurement, 2007, 67(3), 373-393.

[13] Han, K. T., & Hambleton, R. K. 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, 2009.

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

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

[16] Othuon, L. O. A. 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 (1998).

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

[18] González-Romá, V., Hernández, A., & Gómez-Benito, J. 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, (2006), 41(1), 29-53.

[19] Pedrajita, Q J., & Talisayon, V.M. 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, 2009, Vol. 67 (1), 21-43.

[20] Adedoyin, O.O. IRT approach to detect gender biased items in public examinations: A case study from the Botswana junior certificate examination in Mathematics. Educational Research and Reviews 2010, Vol. 5 (7), pp. 385-399.

[21] Fidalgo, Á. M., Ferreres, D. & Muñiz, J. 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, 2004, 73(1), 23-39.