EFFECT OF THE ABILITY DISTRIBUTION SHAPE ON THE GENERALIZED MANTEL-HAENSZEL STATISTICS USED FOR DIF DETECTION

Angel Manuel FIDALGO[1]
Maria Laura QUINTANILLA COBIAN[2]

- ABSTRACT: The main objective of this simulation study was to explore the effect of shape of the θ distribution on the generalized nominal and ordinal Mantel-Haenszel statistics used for detecting DIF in polytomous items. The variables manipulated were: trait (θ), distribution shape (normal, positively skewed, and platykurtic), θ distribution difference between the reference and the focal group (equal and unequal), sample size (500/ 500 and 500/250 examinees in the reference/focal group), and DIF conditions (No DIF, constant and shift-high DIF patterns). The generalized ordinal Mantel-Haenszel statistic was calculated using integer and log-rank scores. The results show: a) a little impact of the θ distribution shape on the performance of all the statistics, and b) the advantages of employing log-rank scores, especially when the items show a shift-high DIF pattern.

- KEYWORDS: Differential item functioning; nonnormal distributions; partial credit model; log-rank scores; polytomous items.

1 Introduction

The Mantel-Haenszel methods constitute one of the most popular nonparametric differential item functioning (DIF) detection procedures. In the case of polytomous items, generalizations of the MH chi-squared statistic χ^2_{MH} have also been used for detecting DIF: the generalized Mantel-Haenszel test - GMH (MANTEL and HAENSZEL, 1959; ZWICK *et al*., 1993a) and the Mantel test (MANTEL, 1963; ZWICK *et al*., 1993a). Fidalgo and Madeira (2008) have showed a unified framework for the analysis of DIF using the generalized Mantel-Haenszel statistic proposed by Landis *at al.* (1978). As is pointed out there, the GMH test and the Mantel test are particular cases of the generalized nominal Mantel-Haenszel ($Q_{GMH(1)}$) statistic and the generalized ordinal Mantel-Haenszel ($Q_{GMH(2)}$) statistic, respectively. In the same article they showed the results of a little simulation study about the effect of the choice of scores assigned to the response variables on the $Q_{GMH(2)}$ statistic. They found that the use of log-rank scores instead of the usual integer scores increased the power of $Q_{GMH(2)}$ for detecting the shift-high. This topic has received little attention given that studies on DIF have routinely employed integer scores

[1] Universidad de Oviedo, Faculdad de Psicología, Plaza Feijoo, s/n, 33003, Oviedo, Asturias, Spain. E-mail: *fidalgo@uniov.es*
[2] Universidad Nacional de Educación a Distancia, Facultad de Psicología, Departamento de Metodología de las Ciencias del Comportamiento, C/ Juan del Rosal 10, 28040, Madrid, Spain. E-mail: *lquintanilla@psi.uned.es*

(ANKENMANN *at al*., 1999; SU and WANG, 2005; WANG and SU, 2004b; ZWICK and THAYER, 1996).

On the other hand, although there is a increased interest in the development of psychometric theories that allow work with nonnormal distributions (SAMEJIMA, 2000; BAZÁN *et al*., 2006) and research about the influence of the nonnormality over the parameter recovery (KIRISCI *et al*., 2001; REISE and YU, 1990; VAN DER OORD *et al*., 2003), there is little research about the effect of the nonnormality on the DIF detection procedures.

Bearing in mind the above, the main goal of the present study is to determine whether the capability of the $Q_{GMH(1)}$ and $Q_{GMH(2)}$ statistics for DIF detection is affected by the shape of the θ distribution and, in the case of $Q_{GMH(2)}$, for the choice of scores assigned to the ordinal variable. With this objective a test was constructed to replicate educational tests that contain both dichotomous and polytomous items.

2 Generalized Mantel-Haenszel statistics

Type Below, we briefly present the generalized Mantel-Haenszel statistics used in this study. The interested reader can find more comprehensive information on these statistics in Fidalgo (2005) and Fidalgo and Madeira (2008). Landis *et al*. (1978) proposed a generalized MH statistic for the analysis of Q: $R \times C$ contingency tables. The data structure for this general contingency table is shown in Table 1.

Table 1 - Data structure in the h_{th} stratum

Factor levels	Response Variable Categories 1	2	⋯	j	⋯	C	Total
1	n_{h11}	n_{h12}	⋯	n_{h1j}	⋯	n_{h1C}	$N_{h1\cdot}$
2	n_{h21}	n_{h22}	⋯	n_{h2j}	⋯	n_{h2C}	$N_{h2\cdot}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	
i	n_{hi1}	n_{hi2}	⋯	n_{hij}	⋯	n_{hiC}	$N_{hi\cdot}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	
R	n_{hR1}	n_{hR2}	⋯	n_{hRj}	⋯	n_{hRC}	$N_{hR\cdot}$
Total	$N_{h\cdot 1}$	$N_{h\cdot 2}$	⋯	$N_{h\cdot j}$	⋯	$N_{h\cdot C}$	N_h

The standard generalized Mantel-Haenszel is defined by Landis *et al*. (1978) as:

$$Q_{GMH} = \left\{\sum_{h=1}^{Q}(\mathbf{n}_h - \mathbf{m}_h)'\mathbf{A}_h'\right\}\left\{\sum_{h=1}^{Q}\mathbf{A}_h\mathbf{V}_h\mathbf{A}_h'\right\}^{-1}\left\{\sum_{h=1}^{Q}\mathbf{A}_h(\mathbf{n}_h - \mathbf{m}_h)\right\} \qquad (1)$$

where $\mathbf{n}_h$, $\mathbf{m}_h$, $\mathbf{V}_h$ and $\mathbf{A}_h$ are, respectively, the vector of observed frequencies, the vector of expected frequencies, the covariances matrix, and a matrix of linear functions defined in accordance with the alternative hypotheses (H_1) of interest. The null hypotheses of no-association will be tested against different H_1: a) general association (both variables are nominal), b) mean score differences (factor is nominal and response ordinal); and c) linear correlation (both variables are ordinal). From Table 1, these vectors and matrices are defined as:

$$\mathbf{n}_h = (n_{h11}, n_{h21}, \ldots, n_{hRC})' \quad (CR \text{ x } 1),$$

$$\mathbf{m}_h = N_{h\cdot\cdot}(\mathbf{p}_{h\cdot *} \otimes \mathbf{p}_{h *\cdot}) \quad (CR \text{ x } 1),$$

$$\mathbf{V}_h = N_{h\cdot\cdot}^2/(N_{h\cdot\cdot}-1)\{(\mathbf{D}_{p_{h\cdot *}} - \mathbf{p}_{h\cdot *}\mathbf{p}_{h\cdot *}^{'})\otimes(\mathbf{D}_{p_{h *\cdot}} - \mathbf{p}_{h *\cdot}\mathbf{p}_{h *\cdot}^{'})\} \quad (CR \text{ x } CR),$$

where $\mathbf{p}_{h*\cdot}$ and $\mathbf{p}_{h\cdot *}$ are, respectively, ($R \times 1$) and ($C \times 1$) vectors with the marginal row proportions ($p_{hi\cdot} = N_{hi\cdot}/ N_{h\cdot\cdot}$) and the marginal column proportions ($p_{h\cdot j} = N_{h\cdot j}/ N_{h\cdot\cdot}$), $\otimes$ denoting the Kronecker product multiplication, $\mathbf{D}_{ph*}$ is a ($C \times C$) diagonal matrix with elements of the vector $\mathbf{p}_{h*}$on its main diagonal, and $\mathbf{D}_{ph*}$ is an ($R \times R$) diagonal matrix with elements of the vector $\mathbf{p}_{h*}$ on its main diagonal.

In this study we will use the generalized MH statistics employed to test the general association ($Q_{GMH(1)}$) and the mean score difference hypotheses ($Q_{GMH(2)}$). To obtain them we should resolve the equation 1 using different matrices $\mathbf{A}_h$ ($\mathbf{A}_h = \mathbf{C}_h \otimes \mathbf{R}_h$). Briefly, these are.

- *$Q_{GMH(1)}$ or the Generalized Nominal MH statistic (GNMH).* Here, $\boldsymbol{R}_h = [\boldsymbol{I}_{R\text{-}1}, -\boldsymbol{J}_{R\text{-}1}]$ and $\boldsymbol{C}_h = [\boldsymbol{I}_{C\text{-}1}, -\boldsymbol{J}_{C\text{-}1}]$, where $\boldsymbol{I}_{R\text{-}1}$ is an (R-1 x R-1) identity matrix, and $\boldsymbol{J}_{R\text{-}1}$ is an (R-1 x 1) vector of ones. Thus, the dimension of $\boldsymbol{R}_h$ will be (R-1 x R). Similarly, $\boldsymbol{I}_{C\text{-}1}$ is an (C-1 x C-1) identity matrix, and $\boldsymbol{J}_{C\text{-}1}$ is a (C-1 x 1) vector of ones. Under H_0, $Q_{GMH(1)}$ follows approximately a chi-squared distribution with degrees of freedom (df) = (R-1)(C-1). When $R = C = 2$, $Q_{GMH(1)}$ is identical to the χ^2_{MH} statistic, except for the lack of the continuity correction. For the special case of 2 factor levels, $Q_{GMH(1)}$ is identical to the generalized Mantel-Haenszel test (GMH) proposed by Mantel and Haenszel (1959).

- *$Q_{GMH(2)}$ or the Generalized Ordinal MH statistic* (*GOMH*). Here $\boldsymbol{R}_h$ is the same as that used in the previous case and $\boldsymbol{C}_h = (c_{h1}, \ldots, c_{hC})$ being a ($1 \times C$) vector, where c_{hj} is an appropriate score reflecting the ordinal nature of the jth category of response for the hth stratum. Under H_0, $Q_{GMH(2)}$ has approximately a chi-squared distribution with df = (R-1). For the special case of 2 factor levels, $Q_{GMH(2)}$ is identical to the extended MH test proposed by Mantel (1963).

Calculation of the $Q_{GMH(2)}$ statistic requires selecting the scores ($\boldsymbol{C}_h$) that will be applied to the response variable to compute the row mean scores $[\bar{y}_{hi} = \sum_{j=1}^{C}(c_{hj}n_{hij} / N_{hi\cdot})]$ used for comparing the factors across strata. In the DIF literature, integer scores are the most common choice (ANKENMANN *et al*., 1999; WANG and SU, 2004; ZWICK and THAYER, 1996), although selection of the values of $\boldsymbol{C}_h$ admits different possibilities as, for example, rank scores.

Table 2 - Frequency tables with the responses given by focal and reference groups to a 4-point item with DIF in the highest category

Ability Level	Group	Response item categories 1	2	3	4
h=1					
	Reference	12	2	0	0
	Focal	11	2	0	0
Total		25	4	0	0
	Log-rank	0.190	-0.629	-1.129	-2.129
h=2					
	Reference	38	14	2	0
	Focal	29	12	0	0
Total		67	26	2	0
	Log-rank	0.304	-0.594	-1.428	-2.428
h=3					
	Reference 3	5	34	8	2
	Focal	37	54	13	0
Total		72	88	21	2
	Log-rank	0.607	-0.186	-1.099	-2.099
h=4					
	Reference 18	61	32	7	
	Focal	18	56	50	4
Total		36	117	82	11
	Log-rank	0.854	0.297	-0.585	-1.585
h=5					
	Reference 5	30	80	14	
	Focal	4	26	75	7
Total		9	56	155	21
	Log-rank	0.963	0.721	-0.159	-1.159
h=6					
	Reference 0	3	30	38	
	Focal	0	7	43	31
Total		0	10	73	69
	Log-rank	1.000	0.934	0.420	-0.508
h=7					
	Reference 0	0	5	30	
	Focal	0	0	5	16
Total		0	0	10	46
	Log-rank	1.000	1.000	0.821	-0.179

Note: Log-rank scores are computed using Equation 2. To avoid zero denominators in Equation 3, in stratums where that can occur, a value of 0.5 was added to each column marginal frequency. In the first stratum, as there are observed ties, the corresponding log-rank scores for breaking ties are: 0.190, -0.629, -1.629 and -1.629.

In the present simulation study, we will employ integer and log-rank scores. From the general contingency table shown in Table 1, we obtain log-rank scores using the equation:

$$c_{hj} = 1 - \sum_{k=1}^{j} \left(\frac{N_{h \cdot k}}{\sum_{m=k}^{C} N_{h \cdot m}} \right) \qquad (2)$$

To avoid zero denominators in Equation 2, in stratums where that can occur, a value of 0.5 was added to each column marginal frequency. In the present study, log-rank scores were calculated in two different ways. In the first of these, ties were broken assigning to ties the average of the values for the corresponding log-rank scores (KOCH *at al.*, 1985). This strategy results in a conservative influence on test statistic. For this reason, in the second way, ties were not broken. In order to show how the log-rank scores were computed, Table 2 shows the frequencies tables with the responses given to a 4-point item that presents a shift-high DIF pattern and the corresponding log-rank scores. These responses were simulated using the partial credit model (PCM) (MASTERS, 1982) described below.

3 Simulation study

3.1 Data generation

Item parameters. An artificial test was constructed having 20 dichotomous items and seven polytomous items with four ordinal response categories. The item parameter values were selected so as to be representative of items found in applied testing setting, and were the same as those used by Chang *et al*. (1996). The generating parameters of the reference group are presented in the Table 3. The modelling of a mixed format test (20 dichotomous items and 4 polytomous items in the matching test) was intended to resemble the currently common practice of combining multiple-choice and constructed-response items in a single test administration. This type of tests have been used in other simulation studies, for example, see Ankenmann *at al*. (1999); Su and Wang (2005) and Zwick *et al*., 1993).

Table 3 - Item parameters. The last three items are the studied items (items 25, 26 and 27)

Dichotomous item parameters			
Item	*a*	*b*	*c*
1	0.741	-2.25	0.15
2	0.861	-2.00	0.15
3	1.162	-1.75	0.15
4	0.638	-1.50	0.15
5	1.000	-1.25	0.15
6	1.000	-1.00	0.15
7	1.162	-0.75	0.15
8	0.638	-0.50	0.15
9	0.741	-0.25	0.15
10	0.861	0.00	0.15
11	1.000	0.00	0.15
12	0.741	0.25	0.15
13	1.162	0.50	0.15
14	0.638	0.75	0.15
15	0.861	1.00	0.15
16	0.638	1.25	0.15
17	0.741	1.50	0.15
18	1.162	1.75	0.15
19	1.000	2.00	0.15
20	0.861	2.25	0.15

Polytomous item parameters			
Item	b_{i1}	b_{i2}	b_{i3}
21	-0.91	-0.93	1.29
22	-1.34	1.72	3.40
23	-1.76	0.09	0.19
24	-2.20	-1.33	-0.48
25	-0.91	0.98	0.21
26	-2.25	-1.80	1.66
27	-0.54	-2.11	0.74

To form DIF items, the b_{ik} parameters of the 25-, 26- and 27-items in Table 3 were changed for the focal group according to the following equations:

$b_{ikF} = b_{ikR} + s,\ k = 1, 2, 3$ (constant DIF pattern)

$b_{ikF} = b_{ikR} + s,\ k = 3$ (shift high DIF pattern)

where *s* is equal to 0.20 and 0.40 under the constant and the shift high DIF pattern, respectively.

θ parameter. Normal, positively skewed and platykurtic θ distributions shapes were used to generate the data. Under all the distribution shapes, the ability of the reference group was a univariate distribution with mean 0 and standard deviation 1. Moreover, two focal groups were simulated: the first had the same ability distribution as the reference group, and the second had a mean one standard deviation below the reference group mean.

Fleishman's (1978) power function $Y = a + b + c\,Z^2 + d\,Z^3$ was used to generate the skewed deviate, were Y is the positively skewed deviate, Z is a standard normal variable, and the rest of the constants were equal to a =0.1736, b =1.1125, c = 0.1736, and d = -0.0503. To generate the platykurtic distribution, the constant values in the Fleishman's power function were set to a = 0.0, b = 1.2210, c = 0.0, and d = -0.0802. Identical values have been used by Kirisci *at al*. (2001).

It should be noted that we obtain strict skewed (skewness = 0.0, kurtosis = 0.0) and platikurtic distributions (skewness = 0.0, kurtosis = -1.0) using the Fleishman's power function only when Z is a normal variable with mean 0 and standard deviation 1. This is the case for the reference group, and the focal group with the same ability distribution as the reference group. In the other focal group, we used in the power function a normal variable with mean -1 and standard deviation 1. Therefore, in this case, we will obtain non-normal distributions with different values of skewness and kurtosis. Figure 1 show the general shapes of the θ distributions used based on 500 θ values.

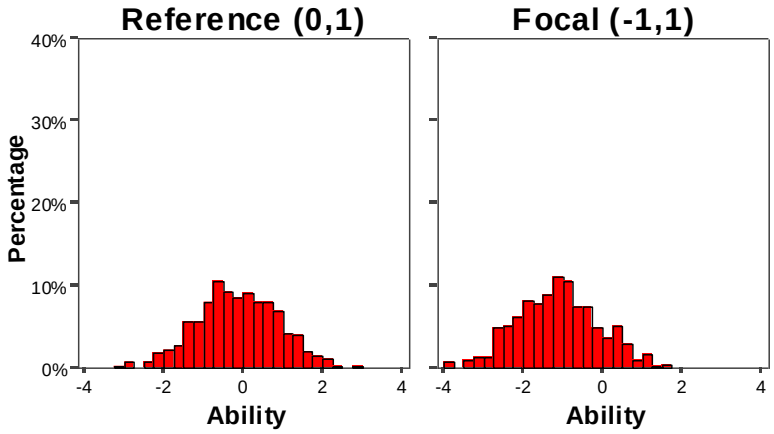

Figure 1a - Shapes of the θ distributions used to generate the data set (Normal).

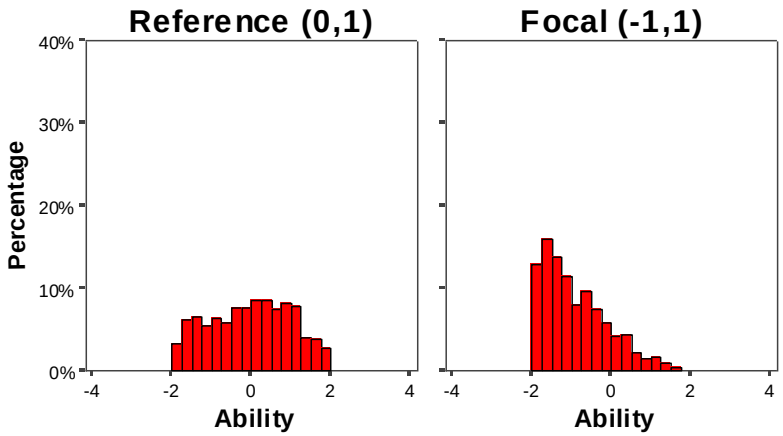

Figure 1b - Shapes of the θ distributions used to generate the data set (platykurtic).

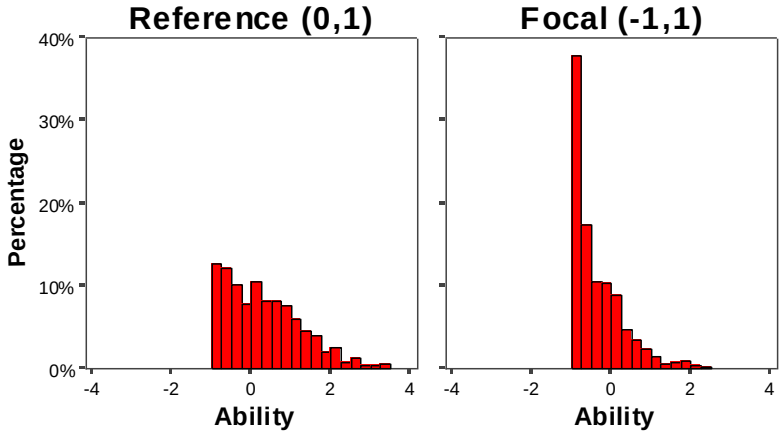

Figure 1c - Shapes of the θ distributions used to generate the data set (Positively skewed).

Models. Datasets were generated for the 20 dichotomous items from a three-parameter logistic IRT model (3PLM) and from the partial credit model (PCM) (MASTERS, 1982) for the 4-point polytomous items. In the 3PLM, the probability of a correct response on item i, for an examinee with latent trait θ, is defined by

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \ ,$$

where a_i is the item discrimination parameter, b_i is the item difficulty parameter, c_i is a pseudo guessing parameter, and D is a scaling factor equal to 1.7.

The polytomous items were generated using the PCM. As it is known, in this model, the probability of scoring x on item i with K categories (from k =1 to K), given θ, is defined by

$$P_{ix}(\theta) = \frac{\exp\left(\sum_{k=1}^{x}(\theta - b_{ik})\right)}{\sum_{h=1}^{K-1}\exp\left(\sum_{k=1}^{h}(\theta - b_{ik})\right)} \ ,$$

where b_{ik} is the kth item difficulty parameter. By definition $\sum_{k=1}^{1}(\theta - b_{ik}) = 0$.

Sample size. We selected the group size pairs based on their proximity to real data. The larger sample size (500/500 examinees in the reference/focal groups) is consistent with those found in practice and has been used in numerous simulation studies (PENFIELD and ALGINA (2003); SU and WANG (2005); WANG and SU (2004b); ZWICK *et al*., 1993). Smaller sample size (500/250) is found in many precalibration stage DIF studies.

3.2 Form of calculating the generalized MH statistics

The MH statistics compare two or more groups conditional on a measure of ability evaluated by the test (the matching variable). The majority of the applications use the total score in the test as an estimation of ability. In such cases, in order to avoid contamination of the matching variable by the items with DIF, it is essential to apply this methodology in two stages (HOLLAND and THAYER, 1985, FIDALGO *et al*., 2000; WANG and SU, 2004a, 2004b). With the aim of comparing the manipulated variables in an optimum situation, we calculated the generalized MH statistics using for calculation of total test score only the non-DIF items (Items 1 to 24; see Table 2). The item under analysis was always included in the matching variable. Furthermore, the software developed to calculate the generalized MH statistics was programmed to automatically exclude those levels of the matching variable in which there was only an examinee ($N_{h\cdot\cdot}$=1). Intervals of one unit in the scale of scores were employed for matching examinees.

3.3 Design

The following factors were manipulated to study the effect on Type I error rate and power of the DIF-detection statistics: θ distribution shape (normal, positively skewed, and platykurtic), θ distribution difference between the reference and the focal group (equal and unequal), sample size (500/ 500 and 500/250 examinees in the reference/focal group), and DIF conditions (No DIF, constant and shift-high DIF patterns).

For each of the 3 (θ distribution shape) × 2 (θ distribution difference between groups) × 2 (Sample size) × 3 (DIF conditions) = 72 cells of the design, 1,000 data sets were generated using the GAUSS program (Aptech Systems, 1993, V.3.1.4). In total, 72 × 1,000 = 72,000 different tests were analysed with the $Q_{GMH(1)}$ and $Q_{GMH(2)}$ statistics (employing integer and log-rank scores) using a modification of the GMDIF program (FIDALGO, 2011).

4 Results

Type I error rates were computed as average number of rejections of the studied items generated under H_0 (Non-DIF) over the 1,000 replications. The Bradley (1978) liberal criterion was used in this study to assess the robustness of the MH statistics. A test fulfils his liberal criterion at $\alpha = 0.05$ if the Type I error rate is between 0.025 and 0.075 ($0.5\alpha \leq n \leq 1.5\alpha$). The proportion of correctly-identified each DIF item in 1,000 data sets was used as a power estimate. The average Type I error rate and power of the generalized MH statistics under each θ distribution shape are given in Table 4 (normal), Table 5 (platykurtic), and Table 6 (positively skewed).

In the case of QGMH(2), the estimations obtained averaging log-rank scores for tied observations are shown in square braked. As it can be seen in Table 4 thought 6, applying log-rank scores as if there were no ties give a more sensitive result than averaging log-rank for ties. Therefore, from now on all mention of the power and Type I error rate will refer to the result of log-rank scores computed without breaking ties.

In our study all three MH statistics yielded Type I error rates for the null- case that fulfilled Bradley's liberal robustness criterion under all the simulated conditions.

As expected, sample size had a great effect on power. The power of all the MH statistics for detecting DIF is very low under the smallest sample size, ranging from .26 to 0.42 (constant DIF) and 0.11 to 0.32 (shift-high DIF), across all θ distribution shapes and between groups ability distributions. In the largest sample size (500/500) the power found ranged from 0.37 to 0.57 (constant DIF) and 0.17 to 0.44 (shift-high DIF), across all the conditions (see Table 4, 5 and 6).

An analysis of tables 3 through 5 reveals that the impact of the θ distribution shape on the power of QGMH(1) and QGMH(2) is small. Normal and platykurtic distributions have very similar results. On the other hand, the positively skewed distribution shows a small increased in power in relation to the other distributions only when the DIF pattern was shift-high. This increase ranged from 2% to 8%. This result is explained for the higher number of examinees in the response category where the DIF is present because of the skewed distribution of θ compared with the normal and platykurtic distributions (see Figure 1).

Table 4 - Average type I error rate and power for the studied items under the normal ability distribution conditions. The estimations in square brackets were obtained averaging log-rank scores for tied observations

Sample size	Between groups DIF condition	Ability distribution	Generalized MH statistics $Q_{GMH(1)}$	$Q_{GMH(2)\text{-integer}}$	$Q_{GMH(2)\text{-Logrank}}$
500/ 500	No DIF	Equal	.047	.055	.053 [.052]
		Unequal	.051	.055	.057 [.054]
	Constant	Equal	.40	.57	.55 [.53]
		Unequal	38	.54	.52 [.49]
	High	Equal	.42	.32	.40 [.38]
		Unequal	.24	.17	.23 [.22]
500/ 250	No DIF	Equal	.049	.051	.047 [.050]
		Unequal	.049	.041	.043 [.044]
	Constant	Equal	.27	.41	.39 [.36]
		Unequal	.28	.41	.39 [.36]
	High	Equal	.29	.21	.27 [.26]
		Unequal	.16	.12	.14 [.13]

Table 5 - Average type I error rate and power for the studied items under the platykurtic ability distribution conditions. The estimations in square brackets were obtained averaging log-rank scores for tied observations

Sample size	Between groups DIF condition	Ability distribution	Generalized MH statistics $Q_{GMH(1)}$	$Q_{GMH(2)\text{-integer}}$	$Q_{GMH(2)\text{-Logrank}}$
500/ 500	No DIF	Equal	.051	.047	.051 [.052]
		Unequal	.042	.043	.046 [.042]
	Constant	Equal	.40	.57	.55 [.52]
		Unequal	.38	.55	.53 [.50]
	High	Equal	.44	.31	.39 [.39]
		Unequal	.25	.17	.21 [.19]
500/ 250	No DIF	Equal	.051	.050	.053 [.051]
		Unequal	.050	.052	.053 [.049]
	Constant	Equal	.28	.42	.39 [.36]
		Unequal	.26	.42	.39 [.37]
	High	Equal	.28	.22	.28 [.26]
		Unequal	.15	.11	.13 [.13]

Table 6 - Average type I error rate and power for the studied items under the skewed ability distribution conditions. The estimations in square brackets were obtained averaging log-rank scores for tied observations

			Generalized MH statistics		
		Between groups			
Sample size	DIF condition	Ability distribution	$Q_{GMH(1)}$	$Q_{GMH(2)\text{-integer}}$	$Q_{GMH(2)\text{-Logrank}}$
500/ 500	No DIF	Equal	.047	.046	.048 [.051]
		Unequal	.050	.051	.050 [.054]
	Constant	Equal	.39	.54	.52 [.50]
		Unequal	.37	.53	.51 [.48]
	High	Equal	.46	.38	.47 [.46]
		Unequal	.31	.23	.28 [.27]
500/ 250	No DIF	Equal	.048	.052	.056 [.051]
		Unequal	.055	.043	.046 [.041]
	Constant	Equal	.27	.40	.39 [.36]
		Unequal	.27	.40	.38 [.36]
	High	Equal	.32	.27	.32 [.30]
		Unequal	.21	.15	.19 [.18]

The results about the DIF patterns corroborate what has been found in the literature comparing QGMH(1) and QGMH(2)-integer: a) QGMH(1) had much more power for detecting the shift-high DIF than QGMH(2), and b) QGMH(2) is more powerful than QGMH(1) for detecting the constant DIF. It should be noted, however, that the difference between QGMH(1) and QGMH(2) for detecting the shift-high DIF was drastically reduced when QGMH(2) was computed using log-rank scores. This finding support the results of Fidalgo and Madeira (2008)

Finally, in general, all the statistics yielded more power under the equal ability distribution conditions than under the unequal conditions. Moreover, as it can be seen in Tables 4 through 6, that this difference was considerably higher under the shift-high DIF pattern (always over the 9% across all the simulated conditions and statistics) than under the constant pattern.

Conclusions

The main topics investigated by the present research was: (a) how the θ distribution shapes affects the performance of both $Q_{GMH(1)}$ and $Q_{GMH(2)}$ statistics; and (b) the effect of log-rank score assigned to the response variable on the $Q_{GMH(2)}$ statistic. Other variables manipulated in the simulation study were the effect of the different θ distribution between groups and the sample size.

One very clear finding is that the statistics applied [$Q_{GMH(1)}$, $Q_{GMH(2)\text{-INTEGER}}$, $Q_{GMH(2)\text{-LOG-RANK}}$] yielded controlled Type I error rates for the null case, fulfilling Bradley's liberal robustness criterion under all the simulated conditions.

A second finding is that the θ distribution shapes have a small impact on the performance of the generalized MH statistics used for DIF detection. In this respect the DIF pattern was the most important factor in order to predict the performance of $Q_{GMH(1)}$ and $Q_{GMH(2)}$. As pointed out in Fidalgo and Madeira (2008), $Q_{GMH(2)}$ increases the statistical power with respect to $Q_{GMH(1)}$ for detecting that the mean responses differ across the factor levels, whereas $Q_{GMH(1)}$ offers the possibility of detecting more complex patterns of association than $Q_{GMH(2)}$. Thus, it is small wonder that the $Q_{GMH(2)}$ statistic yields higher power than $Q_{GMH(1)}$ under the constant DIF pattern whereas $Q_{GMH(1)}$ yields higher power that $Q_{GMH(2)}$ for detecting the shift-high DIF.

A third finding is that, as expected, all the MH statistics, on increasing the sample size, increased their power under all the simulated conditions. On the other hand, the between groups ability distribution had a differential effect depending on the pattern of DIF. In all cases, detection rates decreased when the ability distributions were unequal. However, when the DIF was shift-high, there was a drastic difference in power between the equal and unequal ability distribution conditions.

Finally, from a practical point of view, the most relevant finding is that, as it was hypothesized, the type of score used to compute $Q_{GMH(2)}$ influences its capability for detecting DIF. It should be stressed that the question here is not to choose a score that faithfully describe the "true" distances between ordered categories, but finding a score that allow us to detect the association pattern of interest. Specifically, we have obtained that, when the pattern of DIF is constant, used integer or log-rank scores yield very similar results. However, when the pattern of DIF is shift-high, using log-rank scores offer higher power than the usual integer scores. Moreover, given that the performances of $Q_{GMH(2)\text{-}LOG\text{-}RANK}$ and $Q_{GMH(1)}$ are very similar under the shift-high DIF pattern and $Q_{GMH(2)\text{-}LOG\text{-}RANK}$ his fairly superior to $Q_{GMH(1)}$ for detecting the constant DIF, $Q_{GMH(2)\text{-}LOG\text{-}RANK}$ is recommended if a single method is to be used for detecting the DIF patterns simulated. Unfortunately, in real tests we cannot know, in advance, what type of DIF the items have, and since $Q_{GMH(1)}$ is capable to detect more complex pattern of association than $Q_{GMH(2)}$, $Q_{GMH(1)}$ can be a more realistic option (FIDALGO and BARTRAM, 2010; FIDALGO and SCALON, 2012).

Acknowledgements

We thank reviewers and editors for their comments and suggestions.

FIDALGO, A. M., COBIAN, M. L. Q. Efeito da forma de distribuição de capacidade nas estatísticas generalizadas de mantel-haenszel utilizadas para detecção de DIF. *Rev. Bras. Biom.* Lavras, v.36, n.2, p.438-452, 2018.

- *RESUMO: O objetivo principal deste estudo de simulação foi explorar o efeito da forma da distribuição θ na estatística generalizada nominal e ordinal de Mantel-Haenszel usada para detectar DIF em itens politômicos. As variáveis analisadas foram: forma da distribuição do traço (θ) (normal, enviesada e platicúrtica), diferença na distribuição θ entre a referência e o grupo focal (igual e desigual), tamanho da amostra (500/500 e 500/250 examinadores no grupo referência / focal), e condições DIF (configurações Sem DIF, DIF constante e DIF alto). A estatística ordinal de Mantel-Haenszel foi calculada usando-se os escores de inteiros e log-rank. Os resultados mostram: a) um pequeno impacto da forma de distribuição de θ no desempenho de*

todas as estatísticas e b) as vantagens de empregar pontuações de log-rank, especialmente quando os itens mostram um padrão de DIF com deslocamento elevado.

- *PALAVRAS-CHAVE: Funcionamento diferencial do item, distribuições não normais, modelo de crédito parcial, pontuações log-rank, itens politômicos.*

References

ANKENMANN, R. D.; WITT, E. A.; DUNBAR, S. B. An investigation of the power of the likelihood ratio goodness-of-fit statistic in detecting differential item functioning. *Journal of Educational Measurement*, v.36, n.4, p.277-300, 1999.

APTECH SYSTEMS. *GAUSS (Version 3.1.4) [Computer programming language]*. Aptech Systems, Inc., Maple Valley, WA: USA, 1993.

BAZÁN, J. L.; BRANCO, M. D.; BOLFARINE, H. A skew item response model. *Bayesian Analysis*, v.1, n.4, p.861-892, 2006.

BRADLEY, J. V. Robustness? *The British Journal of Mathematical & Statistical Psychology*, v.33, p.144-152, 1978.

CHANG, H. H.; MAZZEO, J.; ROUSSOS, J. Detecting DIF for politomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement*, v.33, p.333-353, 1996.

FIDALGO, A. M. Mantel-Haenszel methods. In: EVERITT, S; HOWELL, D. C. (Eds.). *Encyclopedia of statistics in behavioral Science*. Chichester: John Wiley & Sons, 2005. v.3, B. p.1120-1126.

FIDALGO, A. M. GMHDIF: A computer program for detecting DIF in dichotomous and polytomous items using generalized Mantel-Haenszel Statistics. *Applied Psychological Measurement*, v.35, n.3, p.247-249, 2011.

FIDALGO, A. M.; BARTRAM, D. A comparison between some generalized mantel-haenszel statistics for detecting DIF in data simulated under the graded response model. *Applied Psychological Measurement*, v.34, n.8, p.600-606, 2010.

FIDALGO, A. M; MADEIRA, J. M. Generalized Mantel-Haenszel methods for differential Item functioning detection. *Educational and Psychological Measurement*, v.68, n.6, p.940-958, 2008.

FIDALGO, A. M.; SCALON, J. D. Using Mantel-Haenszel methods for detecting differential item functioning. *Psicologia-Reflexao e Critica*, v.25, n.1, p.60-68, 2012.

FLEISHMAN, A. I. A method for simulating nonnormal distributions. *Psychometrika*, v.43, p.521-532, 1978.

GRAUBARD, B. I.; KORN, E. L. Choice of column scores for testing independence in ordered 2 x K contingency tables. *Biometrics*, v.43, p.471-476, 1987.

HOLLAND, W. P.; THAYER, D. T. Differential item performance and the Mantel-Haenszel procedure. In: WAINER, H.; BRAUN, H. I. (Eds.). *Test validity*. Hillsdale: LEA, 1988. p.129-145.

KIRISCI, L.; HSU, T.; YU L. Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement*, v.25, p.146-162, 2001.

KOCH, G. G.; SEN, P. K.; AMARA, I. Logrank scores, statistics and tests. In: S. KOTZ AND N. L. JOHNSON (Eds.). *Encyclopedia of statistical sciences.* New York: Wiley. 1985. v.5, p.136-142.

LANDIS, J. R.; HEYMAN, E. R.; KOCH, G. G. Average partial association in three-way contingency tables: A review and discussion of alternative tests. *International Statistical Review*, v.46, p.237-254, 1978.

MANTEL, N. Chi-square tests with one degree of freedom; extension of the Mantel-Haenszel procedure. *Journal of the American Statistical Association*, v.58, p.690-700, 1963.

MANTEL, N.; HAENSZEL, W. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, v.22, p.719-748, 1959.

MASTERS, G. N. A Rasch model for partial credit scoring. *Psychometrika*, v.47, p.149-174, 1982.

REISE, S. P.; YU, J. Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement*, v.27, p.133-144, 1990.

SAMEJIMA, F. Logistic positive exponent family of models: virtue of asymmetric item characteristics curves. *Psychometrika*, v.65, p.319-335, 2000.

SU, Y.-H.; WANG, W. -C. Efficiency of the Mantel, generalized Mantel-Haenszel, and logistic discriminant function analysis methods in detecting differential item functioning in polytomous items. *Applied measurement in education*, v.18, p.313-350, 2005.

VAN DEN OORD, E. J. C. G.; PICKLES, A.; WALDMAN, I. D. Normal variation and abnormality: an empirical study of the liability distributions underlying depression and delinquency. *Journal of Child Psychology and Psychiatry*, v.44, p.180-192, 2003.

WANG, W. -C.; SU, Y.-H. Effect of average signed area between two item characteristic curves and test purification procedures on the DIF detection via the Mantel-Haenszel method. *Applied Measurement in Education*, v.17, p.113-144, 2004a.

WANG, W. -C.; SU, Y.-H. Factors influencing the Mantel and Generalized Mantel-Haenszel methods for the assessment of differential item functioning in polytomous items. *Applied Psychological Measurement*, v.28, p.450-480, 2004b.

ZWICK, R.; DONOGHUE, J. R.; GRIMA, A. Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, v.30, p.233-251, 1993.

Received on 16.10.2016

Approved after revised on 14.07.2017