EDUCARE:
International Journal for Educational Studies, 3(2) 2011

Evaluation of Mantel-Haenszel Statistic for Detecting Differential Item Functioning

Nabeel Abedalaziz

ABSTRACT: The educators have been redefining the goals of instruction and learning to include increased attention to high-level thinking skill. Mantel-Haenszel methods comprise a highly flexible methodology for assessing the degree of association between two categorical variables, whether they are nominal or ordinal, while controlling for other variables. The versatility of Mantel-Haenszel analytical approaches has made them very popular in the assessment of the DIF (Differential Item Functioning) of both dichotomous and polytomous items. The Mantel-Haenszel (M-H) procedure was originally used to Match subjects retrospectively on cancer risk factors in order to study current cancer rates (Mantel & Haenszel, 1959). The terminal objective of the study was to find out the impact of the number of score groups and the inclusion or exclusion of the studied item in forming score groups on estimating αs. Results indicated that: (1) fourth or more score groups yields stable α estimates with Mantel-Haenszel approach; and (2) the inclusion of the studied item is convergent to result in fewer items with significant chi-square values than the exclusion of the studied item in forming score groups. These findings seem to be consistent with the previous researches.

KEY WORDS: Differential Item Functioning, Mantel-Haenszel method, bias, estimating, and inclusion or exclusion of the studied item.

Introduction

In recent years, educators have been redefining the goals of instruction and learning to include increased attention to high-level thinking skill (e.g. National Council of Teaching in Mathematics, 1989). At the same time, educators and psychometricians have been reevaluating how best to assess students' thinking and reasoning skills. Consequently, there has been an increased interest in the use of performance assessments because they have the potential for allowing students to display their solution processes and reasoning. However, evidence is needed to ensure reliable and valid assessments of students' high-level thinking skills. In particular, evidence is needed to ensure that inferences made from performance assessments are equally valid for different subgroups in the population, therefore, the detection of Differential Item Functioning (DIF) is important in addressing issues regarding the quality of the assessments instrument (Wang & Lane, 1996).

Dr. Nabeel Abedalaziz is a Lecturer at the Department of Educational Psychology and Counseling, Faculty of Education UM (University of Malaya), 50603 Kuala Lumpur, Malaysia. He can be reached at: nabeelabdelazeez@yahoo.com and nabilaziz@um.edu.my

DIF refers to differences in item functioning *after* groups have been matched with respect to the ability or attribute that the item purportedly measures. DIF is an *unexpected* difference among groups of examinees who are supposed to be comparable with respect to the attribute measured by the item and the test on which it appears (Dorans & Holland, 1993).

DIF methods therefore assess the test-takers' response patterns to specific test items. DIF occurs when a statistically significant difference is evident in the probability that test-takers from the two distinct groups, who have the same underlying ability on the measured construct, demonstrate differing probabilities of correctly answering the item. As stated, examinees' ability levels are based upon their total scores on the test. As such, the DIF analysis of one specific test item is as independent as possible from the DIF analyses of the other test items (Zumbo, 1999).

To reiterate, a test item is considered to be biased when a dimension on the test is deemed to be irrelevant to the construct that is being measured, placing one group of examinees at a disadvantage in taking a test (Hambleton & Rogers, 1989). Thus, if DIF is not evident for an item, then there is no item bias. Conversely, DIF is required but is not sufficient for item bias. That is, if DIF is apparent, then its presence is not sufficient to declare item bias. An item might show DIF, but not be considered biased if the difference is a result of the actual difference in the groups' ability to respond to the item. If test-takers differed in knowledge, a difference in item responses would be expected. Consequently, a difference in the performance of groups of examinees with different abilities on specific items is not indicative of test bias, but rather of item impact (Schumacher, 2005). But it can be added, that in order to be able to determine whether an item that shows DIF is biased or not, further analysis have to be done (Camilli & Shepard, 1994). It is then of interest to determine whether the differences deepened on differences of ability of the compared groups (not biased) or on the item measuring something else than intended (biased).

On the Mantel-Haenszel (M-H) Procedure

One of the most popular procedures for assessing DIF (Differential Item Functioning) in dichotomous *items* is the Mantel-Haenszel (M-H) procedure. First developed for use in epidemiological research (Mantel & Haenszel, 1959), and later applied to the detection of DIF by P.W. Holland and D.T. Thayer (1986).

The Mantel-Haenszel method works with the item responses for the two groups (referred to in the psychometric literature as the *reference* group and the *focal* group). As described earlier, examinees are first stored into score groups according to total test score, resulting in up to (n + 1) score groups, where n is the number of items in the test. Within the jth score groups, a 2 × 2 table of frequencies is set up:

	Item Score		
	1	0	
Reference Group	A_j	B_j	n_{Rj}
Focal Group	C_j	D_j	n_{Fj}
	m_{1J}	m_{0j}	

A_j, B_j, C_j, and D_j correspond to the number of examinees in the four cells of the 2 × 2 table; n_{Rj}, n_{Fj}, m_{1j}, and m_{1j} are the marginal's. T_j is the number of examinees in the *jth* score groups who attempted the item number investigation. The Mantel-Haenszel test statistic from P.W. Holland and D.T. Thayer (1986) has the form:

$$MHx^2 = \frac{\left(\left|\sum_{j=1}^{s}\left(A_j - E\left(A_j\right)\right)\right| - 1/2\right)^2}{\sum_{j=1}^{s} VAR(A_j)}$$

Where:

$$VAR(A_j) = \frac{n_{Rj} n_{Fj} m_{1j} m_{0j}}{T^2{}_j (T_j - 1)}$$

$$E(A_j) = \frac{n_{Rj} m_{1j}}{T_j}$$

MHx^2 is distributed approximately as a chi-square statistic with one degree of freedom. The term $A_j - E\left(A_j\right)$represents the discrepancy between the observed number of correct responses on the item by Reference group and the expected number. When the observed number is higher than the expected, $A_j > E\left(A_j\right)$, this indicates the potential for DIF in favor of the Reference group, whereas the opposite is true if $A_j < E\left(A_i\right)$. The Log Odds Ratio (α_{MH}) is a measure of association, and $\beta_{MH} =$ Log (α_{MH}) is a signed index. A positive value signifies DIF in favor of the Reference group, and a negative value indicates DIF in favor of the Focal group. If the null hypothesis is true, this quantity is zero.

This statistic has the chi-square distribution with one degree of freedom. Mantel-Haenszel statistics exceeding the tabulated value of the chi-square distribution at a specified level of alpha indicate that item performance in the reference and focal groups over the (n + 1) score groups is consistently different.

Two aspects of special concern to potential user of the M-H technique are: (a) how many score groups to use; and (b) whether or not to include the studied item in the total raw score used to form score groups. J.D. Scheuneman (1979) recommended the use of three to six groups for her chi-square technique for assessing item bias.

P.W. Holland and D.T. Thayer (1986) are recommending a two-step procedure that includes the studied item. This procedure, however, requires a preliminary DIF analysis to purify the matching criterion. Therefore, there is a need to experimental assess how the α indices are affected by the inclusion and exclusion of the studied item in forming score groups.

D.J. Wright (1986) studied the effect of the number of score groups on the delta indices. He found that the fewer the score groups, the greater average of delta indices. Also N.S. Raju, R.K. Bod and V.S. Larsen (1989) determined the effect of number score groups and the inclusion or exclusion of the studied item in forming score groups on estimating αs. They found that four or more groups appear to provide stable α estimate, and the inclusion of the studied item seems to result in fewer items with significant chi-squares than the exclusion of the studied item in forming score groups. In the present study, the researcher applied the same technique.

Purpose and Method:
A. Description of the Test Data and Examinees Samples

The purpose of the study, therefore, is to empirically evaluate the effect on the α indices from the M-H technique of (a) the number of score groups; and (b) the inclusion or exclusion of the studied item in forming score groups.

The samples used in the study were drawn from a data set containing the responses of approximately 1,500 tenth grade students (740 males and 760 females) to a standardized mathematical ability scale. The scale compressed of 60 dichotomous items. The scale was administered in 2009/2010 school year in Malaysia.

B. Procedure

The DIF analysis using the M-H technique was conducted for the Male-Female comparison. In Male-Female comparison, males were treated as the reference group and females as the focal group. For the comparison, 12 different DIF analyses were performed with the M-H technique to assess the effect of the number of score groups and the inclusion or exclusion of the studied item on α estimate.

Using the total raw score on the Mathematical ability scale as the matching criterion, the male and female examinees were separately divided into two mutually exclusive score groups. The first score group was formed by including those examinees whose total raw scores were greater than or equal to 0 and less than or equal to 30. The second score group was similarly formed by including examinees whose total raw scores fell in the closed interval extending from 31 to 60. The two resulting score groups formed the basis for the first DIF analysis.

The second DIF analysis contained the same two score-group cutoffs except that the raw score used for classifying examinees into different score groups did not include the score from the studied item. That is, even though the same cutoff scores (0, 31, 60) were used for forming the two score groups (G_2: two score groups), the total raw score was differently defined for each studied item. The first DIF analysis was performed with the Studied Item Included (SII) in the definition of the total raw score and the second DIF analysis was done with the Studied Item Excluded (SIE) the definition of the total raw score. The next two DIF analyses (separately for SII and SIE) were contained with four score groups using the following cutoffs in forming the score groups: 0, 15, 30, 45, and 60 (G_4).

Similar but separate DIF analysis were also conducted with G_6, G_8, G_{10}, and G_{12} score groups where:

G_6: (six score groups).
G_8: (eight score groups).
G_{10}: (ten score groups).
G_{12}: (twelve score groups).

This procedure resulted in 12 different DIF analyses, with 2 analyses for each of the 6 different numbers of score groups. Finally, the 12 different sets of α estimate were intercorrelated.

Result and Discussion

Table 1 shows the test-score summary for the male and female examinees. The mean of 36.75 for the male group is about 7 raw-score points higher than the mean for the female. The standard deviations are comparable for the two groups. The Kuder-Richardson Formula 20 (KR-20) estimate of reliability is 0.89 for male group and 0.91for female group.

Table 1
Summary Statistic for the Male and Female Examinees

Group	Mean	Standard Deviation	KR-20	Number of Examinees
Male	36.75	7.45	0.89	740
Female	29.75	7.51	0.91	760

Table 2
Proportions (p) Passing the Item and Point-Biserial Correlation (*r*)
for the Male and Female Examinees

	Male		Female			Male		Female	
Item	P	(r)	P	(r)	Item	P	(r)	P	(r)
1	.89	.32	.71	.41	31	.53	.47	.32	.43
2	.82	.42	.71	.45	32	.85	.51	.67	.50
3	.87	.47	.71	.46	33	.71	.50	.50	.44
4	.81	.40	.66	.35	34	.75	.47	.58	.46
5	.74	.52	.54	.48	35	.60	.54	.39	.46
6	.73	.42	.58	.43	36	.60	.61	.36	.45

7	.65	.46	.49	.41	37	.63	.50	.47	.44
8	.65	.41	.40	.46	38	.64	.54	.43	.52
9	.75	.49	.56	.53	39	.47	.45	.28	.42
10	.60	.59	.41	.56	40	.53	.51	.30	.45
11	.72	.47	.48	.39	41	.67	.42	.73	.42
12	.60	.42	.43	.40	42	.56	.52	.70	.43
13	.51	.52	.30	.45	43	.69	.41	.73	.45
14	.43	.41	.32	.38	44	..51	.41	.65	.39
15	.51	.41	.39	.41	45	.48	.46	.60	.37
16	.41	.46	.20	.32	46	.53	.52	.61	.40
17	.70	.52	.51	.45	47	.49	.36	55	.41
18	.71	.36	.61	.47	48	.53	.47	.57	.50
19	.71	.47	.54	.36	49	.50	.45	.64	.60
20	.69	.45	.52	.46	50	.53	.46	.67	.51
21	.60	.46	.45	.38	51	.60	.39	.59	.53
22	.65	.39	.57	.41	52	.46	.41	.64	49
23	.74	.41	.59	.47	53	.50	.51	.68	.53
24	.74	.50	.52	.53	54	.50	.52	.67	.51
25	.67	.47	.49	.44	55	.51	.49	.63	.54
26	.59	.44	.45	.40	56	.72	.43	.69	.49
27	.51	.37	.35	.31	57	.73	.39	.57	.51
28	.89	.46	.79	.47	58	.68	.47	.49	.50
29	.83	.53	.72	.50	59	.67	.51	.60	.50
30	.71	.56	.55	.57	60	.68	.50	.56	.45

Table 2 shows the *p* values and point-biserial correlations for the two groups. The range of p-values for males is from 0.41 to 0.89, whereas the range of p-values for females is from 0.20 to 0.79. The summary data in tables 1 and 2 indicate that a Mathematical ability scale is easier for the Males group than it is for the Females group.

Table 3 shows the correlation for the Male-Female comparison. Also shown in this table are the means and standard deviation for the 12 different sets of α estimate. The values in the diagonal of table 3 are the correlation between the SII estimates of α for the six different numbers of score groups. In all six cases, the correlation between the SII (Studied Item Included) and SIE (Studied Item Excluded) estimates of α is 0.998 indicating that the rank ordering of item α is almost the same whether one includes or excludes the studied item in forming the score group. In terms of the extremely high correlation noted in the diagonal of table 3, it appears that both SII and SIE would yield almost identical results.

Table 3
Means, Standard Deviation, and Intercorrelations of αs Across Different Numbers of Score Groups: Male-Female Comparison

	Numbers of Score Groups					
	G_2	**G_4**	**G_6**	**G_8**	**G_{10}**	**G_{12}**
G_2	(.999)	.983	.979	.980	.977	.966
	.986	(.999)	.998	.998	.998	.998
G_6	.986	.997	(.999)	.998	.999	.997
G_8	.980	.998	.998	(.999)	.999	.998
G_{10}	.980	.998	.998	.999	(.999)	.998
G_{12}	.986	.998	.998	.998	.998	(.999)
SII						
M	1.29	1.05	1.07	1.04	1.04	1.06
SD	0.24	0.23	0.20	0.20	0.20	0.20
SIE						
M	1.35	1.17	1.15	1.13	1.12	1.18
SD	0.24	0.23	0.21	0.21	0.21	0.20

Note: SII = Studied Item Included. SIE = Studied Item Excluded. Numbers in parentheses are correlation between SIE and SII α estimates. M = Mean. SD = Standard Deviation.

The data in the upper right triangle of the matrix in table 3 (the data above the main diagonal) show the intercorrelations between α estimates from different numbers of score groups with the Studied Item Included (SII) in the formation of score groups. These correlations are quite high, with the lowest correlation being 0.966 between G_2 score groups and G_{12} score groups. It should be noted, however, that α estimates with G_2 score groups generally correlate somewhat lower with the α estimates based on G_4, G_6, G_8, G_{10}, and G_{12} score groups. The intercorrelations among α estimate from G_4 G_6, G_8, G_{10}, and G_{12} score groups were 0.997 or better. These results seem to imply that having as few as G_2 score groups may not be "optimal" for estimates αs with SII; alternately, $G_{4,}$ G_6 or more score groups appear to yield highly comparable α estimates.

The means and standard deviations of α estimate with SII also appear to bear out this conclusion. The means α for G_2 score groups is 1.29 whereas the means for G_4 or more score groups vary between 1.10 and 1.04. The standard deviations of α estimates for G_2 score groups is 0.24, and it is only slightly higher than the standard deviations of the α estimates for G_4 or more score groups. The intercorrelations in the upper right triangle of the Table 3, along with the means of the α estimates with SII seem to imply that G_4 or more score groups would yield stable α estimates with the M-H technique; setting number of score groups at G_2 does not appear to be "optimal" for the M-H technique with SII.

The intercorrelations in the lower left triangle of the matrix in the table 3 are for the six different numbers of score groups examined with the Studied Item Excluded (SIE). Again, the correlations are quite high, with the number of score groups set at G_2 doing slightly less well (in terms of the magnitude of the observed correlations) than the score groups set at $G_{4,}$ G_6, G_8, G_{10}, and G_{12}. The means of the α estimates with SII also confirm this trend. The means α for G_2 score groups is 1.29, whereas for G_4 or more score groups, the means vary between 1.04 and 1.07. This general trend for SIE is very similar to the trend observed above for SIE.

Table 4 shows the number of items with significant chi squares by number of score groups. The data in this table are presented separately by significance level (.05 and .01) and Male-Female comparison. Two trends seem to characterize the data in this table. *First*, the number of items with significant chi-squares is greater in G_2 score groups. It appears that more items are likely to be identified as revealing DIF in G_2 groups. For example, there are 44 items with significant chi-squares (at the .05 level of significance) for G_2 score groups and only 23 items with significant chi-squares for G_4 score groups and only 26 items for G_6, G_8, G_{10} score groups with SII. This trend appears to be stable across the .05 and .01 levels of significance and across the Male-Female (gender) comparisons. *Second*, including the studied item is likely to yield slightly fewer items with significant chi-squares than excluding the studied item in forming score groups. This trend also appears to be quite stable across significance levels and gender comparisons with one or two minor exceptions. It should be noted, however, that as the number of score group increases, the difference between SII and SIE becomes less pronounced.

Table 4
Numbers of Item with Significant Chi–Squares Values

Number Score Groups	Male Versus Female			
	.05 Level		.01 Level	
	SII	SIE	SII	SIE
G_2	44	48	40	45
G_4	23	27	23	24
G_6	26	27	27	22
G_8	26	25	15	28
G_{10}	26	26	17	20
G_{12}	15	16	18	19

Note: SII = Studied Item Included. SIE = Studied Item Excluded.

Table 5 shows the percentage overlap across score groups for items whose chi-squares are significant at the .05 level, separately for SII and SIE and gender comparisons. For example, of the 44 items identified as revealing DIF with G_2 score groups (see table 4), 22 items or 50% were also identified as revealing DIF with G_4 in the comparisons with SII.

Table 5
Percentage Overlap across Score Group for Significant Item (p < .05)

Score-Groups Comparison	Male Versus Female	
	SII	SIE
G_2 Versus G_4	.50	.50
G_4 Versus G_6	.49	.48
G_6 Versus G_8	.84	.99
G_8 Versus G_{10}	.89	.96
G_{10} Versus G_{12}	.87	.90

The percentage overlap of statistically significant items for the G_2 Versus G_4 comparison is .50 for both SII and SIE. For other comparisons, the percentage overlap is .84 or better. In summary, there is substantially greater consistency in items of which items are being identified as revealing DIF with G_4 or more score groups than with G_2 score groups. The percentage overlap is about the same for SII and SIE, with a slightly higher percentage for SIE.

Conclusion

In conclusion, fourth or more score groups yields stable α estimate with Mantel-Haenszel approach. The inclusion of the studied item is convergent to result in fewer items with significant chi-square values than the exclusion of the studied item in forming score groups. These findings seem to be consistent with the previous researches (Wright, 1986; and Raju, Bod & Larsen, 1989).

References

Camilli, G. & L. Shepard. (1994). *Methods for Identifying Biased Test Items*. California: Sage Publication.

Dorans, N.J. & P.W. Holland. (1993). "DIF Detection and Description: Mantel-Haenszel and Standardization" in P.W. Holland & H. Wainer [eds]. *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., pp.35-66.

Hambleton, R.K. & H.J. Rogers. (1989). "Detection Potentially Biased Test Items: Comparison of IRT Areas and Mantel-Haenszel Methods" in *Applied Measurement in Education*, 2, pp.313-334.

Holland, P.W. & D.T. Thayer. (1986). "Differential Item Performance and the Mantel-Haenszel Procedure". *Paper* presented at the meeting American Educational Research Association.

Holland, P.W. & D.T. Thayer. (1988). "Differential Item Performance and Mantel-Haenszel Procedure" in H. Wainer & H.I. Braum [eds]. *Test Validity*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., pp.129-145.

Mantel, N. & W. Haenszel. (1959). "Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease" in *Journal of the National Cancer Institute*, 22, pp.719-748.

NRC [National Research Council]. (1989). *Every Body Counts.* Washington, DC: National Academy of Science.

Raju, N.S., R.K. Bod & V.S. Larsen. (1989). "An Empirical Assessment of the Mantel-Haenszel Statistic for Studying Differential Item Performance" in *Applied Measurement in Education*, 2(1), pp.1-13.

Scheuneman, J.D. (1979). "A Method for Assessing Bias in Test Items" in *Journal of Educational Measurement*, 16, pp.143-152.

Schumacher, R. (2005). "Test Bias and Differential Item Functioning" in http://www.appliedmeasurementassociates.com.pdf [Accessed at Kuala Lumpur, Malaysia: 18 November 2010].

Wang, N. & S. Lane. (1996). "Detection of Gender-Related Differential Item Functioning in a Mathematical Performance Assessment" in *Applied Measurement*, 12(2).

Wright, D.J. (1986). "An Empirical Comparison of the Mantel-Haenszel and Standardization Methods of Detecting Differential Item Performance" in *Statistical Report*, No.SR-86-99.

Zumbo, B.D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Like(Ordinal) Item Scores*. Ottawa, Canada: Directorate of Human Resources Research and Evaluation.