The Performance of Mantel-Haenszel Procedures in the Identification of DIF Items

Muhammad Naveed Khalid[1]

Abstract

Item bias is a major threat to measurement validity. This research examined the Type I error and power of the MH test by varying the magnitude of DIF, test score (matching criterion) purification types (single-stage, two-stage, and iterative), test length, and sample size on robustness and power of Mantel-Haenszel (MH) DIF detection procedures. Data was simulated under the one-parameter logistic (1PL) model. In the 20% DIF item conditions the two MH procedures are robust and have sufficient power, but in the 40% DIF item conditions robustness violation and insufficient powers occur. The influence of test length on power is rather modest. On the other hand, test score purification improves power, but the size of their effects is much larger in the 40% DIF item conditions than in the 20% DIF item conditions.

Key Words: Differential item functioning, Power, Type 1 error rate, Purification, Mantel-Haenszel

Introduction

Item response theory (IRT) is a body of related psychometric theory that provides a foundation for scaling persons and items based on responses to assessment items. Item response theory relates characteristics of items (item parameters) and characteristics of individuals (latent traits) to responses (Lord & Novick, 1968). Statistical topics related to IRT are estimating parameters, ascertaining how well data fits a model, and investigating the psychometric properties of assessments. IRT is widely applied in the field of educational and psychological testing in a number of applications such as the evaluation of the reliability and validity of tests, optimal item selection, computerized adaptive testing, developing and refining exams, maintaining banks of items for exams, and equating for the difficulties of successive versions of exams (Glas, 1998). However, above mentioned applications assume that the IRT models used hold. Therefore, over the course

[1] University of Cambridge, khalid.m@cambridgeesol..org

of the past two decades the topic of model-fit has become of more and more interest to test developers and measurement practitioners. The presence of misfitting items may potentially threaten the realization of advantages of IRT models. Among the problems and concerns raised over the years with regard to IRT model fit, DIF is one of the most important.

The psychometrically accepted definition of DIF is that an item shows DIF if individuals having the same ability, but from different groups, do not have the same probability of getting the item correct (Kelderman & Macready, 1990). This requires that individuals in the groups being compared are of comparability ability, as measured by a test or other means. If the item performance of two unmatched groups of examinees is being compared, the result may be measure of item impact rather than of DIF (Holland & Thayer, 1988). Several statistical DIF detection methods have emerged (Holland & Thayer, 1988; Shealy & Stout, 1993; Swaminathan & Rogers, 1990; Thissen, Steinberg, & Wainer, 1988; Kelderman & Macready, 1990), and many reviews of DIF methods are provided in the literature (e.g., Camilli & Shepard, 1994; Holland & Wainer, 1993; Millsap & Everson, 1993; Roussos & Stout, 2004). Most of the techniques for the detection of DIF that have been proposed are based on evaluation of differences in response probabilities between groups, conditional on some measure of ability or proficiency.

One of the most used methods is the Mantel-Haenszel (MH) procedure proposed by Holland and Thayer (1988) where the respondent's unweighted sum score is used as a proxy for proficiency, and DIF is evaluated by testing whether the response probabilities, given the unweighted sum scores, differ among the groups. This procedure is particularly attractive because it can be used with fewer examinees, is easy to program and is relatively inexpensive in terms of computer time. The MH procedure provides an estimate of DIF effect size (the Mantel-Haenszel common odds ratio estimator) and also provides a chi-square test of the null DIF hypothesis (the Mantel-Haenszel chi-square statistic) that the odds of a correct response to a given item are equal for members of the focal and reference groups at equal ability level. Usually, the focal and reference group members are matched on total test score. The matching criterion must be a valid and reliable measure, and should not be contaminated by DIF-items. Therefore, two-stage or iterative DIF detection procedures, which purify the test score by eliminating DIF-items at each stage, are recommended (Lord, 1980). Some iterative procedures have been proposed, some based on the analysis of contingency tables such as the iterative logit method (Fidalgo, Mellenbergh & Muñiz, 1998, 2000) or a two-stage version of MH proposed by Holland and Thayer (1988), non-IRT-based DIF methods (Clauser, Mazor, & Hambleton, 1993; French & Maller, 2007; Hidalgo-Montesinos & Gómez-Benito, 2003; Holland & Thayer, 1988; Miller & Oshima, 1992; Navas-Ara & Gómez-Benito, 2002; Wang & Su, 2004a, 2004b, 2010) and others using item response theory (IRT) models (Cancel & Drasgow, 1988; Lautenschlager, Flaherty & Park, 1994; Lord, 1980). However, if tests have many DIF items, then DIF contamination cannot be completely eliminated by scale purification procedures.

In general, these studies showed that iterative procedures resulted in better DIF-item detection than the corresponding noniterative procedures. Moreover, it is expected that the percentage of DIF items does not affect the DIF detection rates to a large extent if Holland and Thayer's (1988) two-stage procedure is used. Thus, Rogers and Swaminathan (1993) reported that the percentage of items with DIF in the test (0% and 15%) did not affect the MH results; and Miller and Oshima (1992) reported that the MH two-stage procedure did not have substantial impact on DIF detection when the percentage of DIF-items (5% or 10%) and the size of DIF were small. Finally, Narayanan and Swaminathan (1994, 1996) reported an overall decrease of approximately 1% to 5% of correctly identified DIF items as the proportion of DIF items increased from 10% to 20%. In all of these simulation studies the results might be influenced by the test length (40 to 80 items), which implies that test had many unbiased items. Moreover, because total test score is used as the criterion variable for grouping examinees, a more reliable test score (longer test length) may result in improved performance of the MH procedure.

However, Rogers and Swaminathan (1993) showed that test length had no significant influence on the power of the MH procedure for DIF detection, but only long tests were used (40- and 80-items). Uttaro and Millsap (1994) used both short (20 items) and moderate (40 items) test lengths, but DIF was presented only in the studied item. For the

20-item test, the MH procedure gave inflated Type I error rate when the groups differed in ability distributions. However, the inflation in the Type I error rate disappeared entirely in the 40-item test. Moreover, test length generally had little effect on the detection rates in both the 20- and 40 item tests. In the present study short (20 items), moderate (40 items) and long tests (60 items) with different percentages of DIF items in the test were used.

As discussed above, the percentage of biased items and test length effect the power of MH procedure for detecting DIF items because these factors determine the accuracy of the matching variable (total test score). Clauser, Mazor and Hambleton (1991) investigated the influence of the matching variable on the effectiveness of the MH procedure for DIF detection. They found that if the original set of items (total test) was regrouped and DIF was investigated within separate subtests, the items changed their DIF status and the number of items identified as DIF increased as the test length decreased. These results, however, might be explained by changes in the dimensionality of the subtests instead of changes in the length of the subtests (Mazor, Hambleton, & Clauser, 1998).

The main goal of the present study was to examine, using simulated data, the influence of percentage of DIF items, test length and purification type (single stage, two stage and iterative) on the power and Type I error rate of the MH procedure for DIF detection. This article is organized as follows. First, a concise frame work of the MH procedure is sketched for the identification of DIF items. Next, framework of study is presented. Then a number of simulation studies of the Type 1 error rate and power are presented. Finally, some conclusions are drawn, and some suggestions for further research are given.

Mantel-Haenszel Procedure

In the MH procedure of Holland and Thayer (1988), which is widely used by testing companies for DIF screening, a 2×2×K table of examinee data is constructed based on item performance (right or wrong), group membership (the *focal group*, which is of primary interest, or the *reference group*), and score on an overall proficiency measure with K levels, used to match examinees. The two examinee groups are then compared in terms of their odds of answering the item correctly, conditional on the proficiency measure. The odds ratio is assumed to be constant over all levels of the proficiency measure.

Assume that there are T_K examinees at the *kth* level of the matching variable. Of these, n_{RK} are in the reference group and n_{FK} are in the focal group. Of the n_{RK} reference group members, A_K answered the studied item correctly while B_K did not. Similarly, C_K of the n_{FK} matched focal group members answered the studied item correctly, whereas D_K did not. The MH measure of DIF can then be defined as

$$MH\ D-DIF = -2.35\ln(\hat{\alpha}_{MH}) \tag{1}$$

where $\hat{\alpha}_{MH}$ is the Mantel–Haenszel (1959) conditional odds-ratio estimator given by

$$\hat{\alpha}_{MH} = \frac{\sum_K A_K D_K / T_K}{\sum_K B_K C_K / T_K} \tag{2}$$

In Equation (1), the transformation of $\hat{\alpha}_{MH}$ places *MH*D-*DIF* (which stands for "Mantel–Haenszel delta difference") on the ETS delta scale of item difficulty (Holland & Thayer, 1985). The effect of the minus sign is to make *MH*D-*DIF* negative when the item is more difficult for members of the focal group than it is for comparable members of the reference group. The Mantel-Haenszel ($\hat{\alpha}_{MH}$) chi-square test with one degree of freedom provides an approximation to the uniformly most powerful unbiased test of the null hypothesis of no DIF (common odds ratio equal to one) versus the hypothesis of constant DIF (common odds ratio not equal to one). Rejection of the null hypothesis suggests that item performance and group membership are associated, conditional on the matching variable.

In making decisions about whether to discard items or flag them for review, however, testing companies may rely instead on categorical ratings of the severity of DIF. Several testing companies have adopted a system developed by ETS for categorizing the severity of DIF based on both the magnitude of the DIF index and the statistical significance of the results (see Zieky, 1993). According to the original version of this classification scheme, a

"C" categorization, which represents moderate to large DIF, requires that the absolute value of *MH* D-*DIF* be at least 1.5 and be significantly greater than 1 (at α = .05). A "B" categorization, which indicates slight to moderate DIF, requires that *MH* D-*DIF* be significantly different from zero (at α = .05) and that the absolute value of *MH* D-*DIF* be at least 1, but not large enough to satisfy the requirements for a C item. Items that do not meet the requirements for either the B or C categories are labeled "A" items, which are considered to have negligible DIF. Items that fall in the C category are subjected to further scrutiny and may be eliminated from tests. For most purposes, it is useful to distinguish between negative DIF (DIF against the focal group, by convention) and positive DIF (DIF against the reference group). This distinction yields a total of five DIF classifications: C–, B–, A, B+, and C+.

Simulation Design

The item parameters used in the simulation were generated under the one parameter logistic model (1PLM). For given examinee and item parameter values, the subject's probability of giving a correct response which is function of the difficulty of item, β_i, and the examinee's unidimensional ability, θ, under the Rasch model is:

$$P_i(\theta) = \frac{\exp(\theta - \beta_i)}{1 + \exp(\theta - \beta_i)} \qquad (3)$$

Item difficulty parameters were matched to the θ distribution and distributed normally (0, 1). The ability distribution of the reference and focal groups were normally distributed with mean 0 and standard deviation 1. A number was randomly selected from a uniform distribution on the interval (0, 1). If this number was smaller than the probability of a correct response, the item was scored 1; otherwise the item was scored 0. Using this simulation model 100, 400 and 1000 reference group examinee response vectors were generated. The same number of subject response vectors was generated for the focal group. DIF was introduced by changing the item difficulty parameters in the focal group, so that each DIF item was more difficult for the focal group. Three variables were manipulated: amount of DIF (0.5, 1.0), test length (20, 40), Percentage of DIF items (20, 40) and purification type. Per condition of the design 1000 replications were performed. Each data set was analyzed using three purification types (single stage [MH-1], two-stage [MH-2], and iterative [MH-I]) in computing the MH chi-square statistic.

Results

The estimated Type I error rates at nominal level α = .05 are reported in Table 1. The columns labeled K, δ and N denote test length, effect size and sample size, respectively. The values beneath 20% and 40% show the Type I error rate in the presence of DIF items. The proportion of times a non DIF item is incorrectly flagged is an estimate of the MH chi-square statistic's Type I error rate. A test is said to be robust if its Type I error rate is near the nominal significance level α. Bradley (1978) formulated a strict and a liberal criterion of robustness. A test fulfils his liberal criterion at α = .05 if the Type I error rate is between .025 and .075.

The estimated Type I error rates at nominal level α = .05 are reported in Table 1. The table shows that for 20% DIF items in a 10-, 20-, or 40-item test each MH-I procedures fulfils Bradley's liberal criterion. Error rate inflates for small sample size irrespective of purification procedure. For 40% DIF items in a test, the two-stage and iterative MH procedures are robust in majority of the combinations, but the single stage MH procedure is not robust; and the main effects of test length, sample size were also observed. Over all, Type I error rate is inflated for two stage and one stage MH procedures except for the iterative MH procedure.

The proportion of times the DIF items was flagged was used as an estimate of the power of the MH chi-square statistic. Cohen (1988) considers a power above .80 to be sufficient at significance level α = .05. The estimated powers at α = .05 are reported in Table 2. The table shows that for the 20% DIF item conditions the power is sufficient, i.e., above .80, in each of the studied cases (10-, 20-, and 40-item test and MH-l, MH-2, and MH-I procedures except for small sample (100) and small test). For 40% DIF items in a 10-item test the power is not sufficient for one stage and two stage MH procedures especially in small sample ; and for 20 and 40-item test the power is sufficient for the two-stage and iterative procedures, but not sufficient for the single-stage procedure.

The effects of Amount of DIF, Test Length, and Purification Type on robustness and power of the MH procedure were studied. Item responses were simulated for focal and reference group subjects, where the two groups have equal ability distributions. The results showed that test length, sample size and purification type have effects, but that the size of the effects differs between tests containing 20% DIF items and tests containing 40% DIF items.

Table 1: Error rates by test length, effect size, sample size and purification type

			Purification Types					
			MH-1		MH-2		MH-I	
K	δ	N	20%	40%	20%	40%	20%	40%
10	0.5	100	0.08	0.26	0.07	0.16	0.06	0.13
		400	0.07	0.20	0.06	0.14	0.05	0.12
		1000	0.06	0.18	0.06	0.10	0.05	0.10
	1.0	100	0.12	0.30	0.09	0.16	0.10	0.16
		400	0.10	0.24	0.07	0.12	0.08	0.12
		1000	0.07	0.18	0.06	0.11	0.07	0.11
20	0.5	100	0.08	0.11	0.07	0.08	0.06	0.07
		400	0.05	0.10	0.05	0.07	0.05	0.06
		1000	0.05	0.10	0.06	0.06	0.05	0.05
	1.0	100	0.10	0.18	0.09	0.16	0.07	0.07
		400	0.08	0.16	0.08	0.08	0.05	0.06
		1000	0.07	0.14	0.06	0.07	0.05	0.05
40	0.5	100	0.13	0.15	0.12	0.14	0.10	0.11
		400	0.07	0.12	0.06	0.10	0.05	0.06
		1000	0.05	0.10	0.05	0.07	0.04	0.05
	1.0	100	0.12	0.18	0.14	0.11	0.06	0.08
		400	0.07	0.17	0.06	0.10	0.05	0.05
		1000	0.06	0.15	0.06	0.08	0.04	0.05

Conclusions

In the 20% DIF item conditions, the two stage and iterative MH procedures are robust and have sufficient power at the 5% significance level. The power increases with test length, but the influence of test length on power is rather large: going from the 10-item to the 20-item test and from the 20-item to the 40-item test the increase in power is most of time greater than .08. The application of the two-stage procedure yields higher power than the single-stage procedure, and the iterative procedure yields higher power than the two-stage procedure, but the effect of purification type is also rather small: going from the single-stage to the two-stage procedure and from the two-stage to the iterative procedure the power increase is always less then .05.

Table 2: Power by test length, effect size, sample size and purification type

			Purification Types					
			MH-1		MH-2		MH-I	
K	δ	N	20%	40%	20%	40%	20%	40%
10	0.5	100	0.28	0.17	0.30	0.19	0.34	0.19
		400	0.75	0.42	0.81	0.46	0.85	0.52
		1000	0.90	0.53	0.97	0.58	1.00	0.63
	1.0	100	0.69	0.31	0.72	0.36	0.77	0.40
		400	0.92	0.37	0.96	0.41	1.00	0.45
		1000	0.94	0.29	0.97	0.31	1.00	0.37
20	0.5	100	0.33	0.30	0.37	0.35	0.40	0.39
		400	0.79	0.72	0.80	0.79	0.84	0.82
		1000	0.93	0.90	0.97	0.95	0.99	0.99
	1.0	100	0.82	0.79	0.85	0.83	0.89	0.87
		400	0.96	0.94	0.98	0.97	1.00	1.00
		1000	0.98	0.96	1.00	0.98	1.00	1.00
40	0.5	100	0.42	0.40	0.50	0.44	0.52	0.48
		400	0.81	0.78	0.85	0.83	0.87	0.87
		1000	1.00	0.98	1.00	0.99	1.00	1.00
	1.0	100	0.88	0.83	0.90	0.86	0.92	0.89
		400	1.00	1.00	1.00	1.00	1.00	1.00
		1000	1.00	1.00	1.00	1.00	1.00	1.00

The results of the 40% DIF item conditions were different. They show some cases of robustness violation and insufficient power. The 10-item test having 40% DIF items has inflated Type I error rate and insufficient power. For 20-item test, two stage and iterative procedure found to be more robust. Similarly, inflated Type I error rate and insufficient power when the single-stage procedure is applied was found for 40-item test especially for small sample, but it is robust and has sufficient power when the two-stage and iterative procedures are applied. Purification Type and Amount of DIF items have substantial influence on power. The application of the two-stage procedure increases power over the single-stage procedure and the iterative procedure increases power over the two stage procedure. The power decreases with percentage of DIF item, the decrease in power is more substantial for MH-l rather than for MH-2, and MH-I, respectively

As seen in results, the iterative procedure is more useful than the two-stage procedure when the number of biased items in the test is large (40%). However, when the number of biased items is small (20%) both the iterative and the two-stage procedures show a similar power and their Type I error rates are within the expected limits. In sum, it is recommended to apply the MH procedure in two stages or iteratively. In practical applications, tests may contain large number of DIF-items. For example, Miller and Linn (1988) reported 20% to 40% DIF-items in tests which were used to evaluate instructional program effects. For test with a large number of DIF items, the iterative MH-procedure is the preferred option.

References

Bradley, J.V. (1978). Robustness? *The British Journal of Mathematical & Statistical Psychology*, 31, 144-152.

Camilli, G. & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Candell, G. L. & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement, 12,* 253-260.

Clauser, B., Mazor, K. M. & Hambleton, R. K. (1991). Influence of the criterion variable on the identification of differential item functioning test item using the Mantel-Haenszel statistic. *Applied Psychological Measurement*, *15*, 353-359.

Clauser, B. E., Mazor, K. M., & Hambleton, R. K. (1993). The effects of purification of the matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education, 6,* 269-279.

Cohen, J. (1988, 2nd ed.). *Statistical power analysis for the behavioural sciences*. Hillsdale, NJ: Erlbaum.

Fidalgo, A. M., Mellenbergh, G.J. & Muñiz, J. (1998). Comparación del procedimiento Mantel-Haenszel frente a los modelos loglineales en la detección del funcionamiento diferencial de los ítems [Comparison of the Mantel-Haenszel procedure versus the loglinear models for detecting differential item functioning]. *Psicothema, 10*, 219-228.

Fidalgo, A. M., Mellenbergh, G.J. & Muñiz, J. (2000). Identification of nonuniform differential item functioning using a variation of the Mantel-Haenszel procedure and the iterative logit method. *Revista Electrónica de Metodología.*

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for differential item functioning detection. *Educational and Psychological Measurement, 67*, 373-393.

Glas, C. A. W. (1998). Detection of differential item functioning using Lagrange multiplier tests. *Statistica Sinica, 8*, 647-667.

Hidalgo-Montesinos, M.D., *&* Gómez-Benito, J. *(*2003*).* Test purification and the evaluation of differential item functioning with multinomial logistic regression. *European Journal of Psychological Assessment*, *19*, 1-11.

Holland, P. W. & Thayer, D. T. (1985). *An alternative definition of the ETS delta scale of item difficulty* (ETS Research Report No. 85 - 43). Princeton, NJ: Educational Testing Service.

Holland, W. P. & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp.129-145). Hillsdale, NJ: Erlbaum.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Kelderman, H., & Macready, G. B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. *Journal of Educational Measurement,* 27, 307–327.

Lautenschlager, G. J., Flaherty, V. L. & Park, D. G. (1994). IRT differential item functioning: An examination of ability scale purifications. *Educational and Psychological Measurement*, *54*, 21-31.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores. Reading* M.A: Addison-Wesley.

Mazor, K.M., Hambleton, R.K., & Clauser, B.E. (1998). Multidimensional DIF analyses: The effects of matching on unidimensional subtest scores. *Applied Psychological Measurement, 22* (4): 357-367.

Miller, M. D. & Oshima, T. C. (1992). Effect of sample size, number of biased items, and magnitude of bias on a two-stage item bias estimation method. *Applied Psychological Measurement*, *16*, 381-388.

Millsap, R. E. & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*, 297-334

Navas-Ara, M. J., & Gómez-Benito, J. (2002). Effects of ability scale purification on identification of DIF. *European Journal of Psychological Assessment*, *18*, 9-15.

Narayanan, P. & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement*, *18*, 315-328.

Narayanan, P. & Swaminathan, H. (1996). Identification of items that show no uniform DIF. *Applied Psychological Measurement*, *20*, 257-274.

Roussos, L. A., & Stout, W. F. (2004). Differential item functioning analysis. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 107-116). Thousand Oaks, CA: Sage.

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 58*, 159-194.

Swaminathan, H., & Rogers, H.J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*, 361-370.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147-169). Hillsdale, NJ: Lawrence Erlbaum.

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

Wang, W.-C., & Su, Y.-H. (2004a). Effects 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*, *17*, 113-144.

Wang, W.-C., & Su, Y.-H. (2004b). Factors Influencing the Mantel and generalized Mantel-Haenszel methods for the assessment of differential item functioning in polytomous items. *Applied Psychological Measurement*, *28*, 450-480.

Wang, W.-C., & Su, Y.-H. (2010). MIMIC Methods for Assessing Differential Item Functioning in Polytomous Items. *Applied Psychological Measurement, 34*, 166-180.

Zwick, R., Thayer, D. T. & Wingersky, M. (1993). *A simulation study of methods for assessing differential item functioning in computer-adaptive tests.* (ETS Research Report 93-11). Princeton, NJ: Educationl Testing Service.