Revista Electrónica de Metodología Aplicada
2000, Vol. 5 nº 1, pp. 13-24

IDENTIFICATION OF NONUNIFORM DIFFERENTIAL ITEM FUNCTIONING USING A VARIATION OF THE MANTEL-HAENSZEL PROCEDURE AND THE ITERATIVE LOGIT METHOD.

Angel M. Fidalgo*
Gideon J. Mellenberg**
José Muñiz*

***University of Oviedo**
****University of Amsterdam**
e-mail: fidalgo@correo.uniovi.es

ABSTRACT.

The standard Mantel-Haenszel (MH) procedure, a simple modification of the MH procedure (reanalyzing the data separately for high- and low-performing groups), and the iterative logit method were compared with respect to their robustness and power to detect symmetric nonuniform diferential item functioning (DIF). Data for a 79-item test with 11% of DIF items were simulated using a three parameter logistic model for focal and reference groups. The ability distributions of both groups were normally distributed. The factors manipulated were sample size (200 and 1,000 examinees per group) and DIF-size (.25 and .50). As expected, results show that sample size and DIF size improve power, although the power is rather low at each of the factor levels. The MH variation yields higher power than the standard MH procedure and the iterative logit method for each cell of the design, but it was not robust in the N = 1,000 sample size condition.

Key words: Differential item functioning, item bias, iterative logit method, logit models, Mantel-Haenszel chi-square statistic, nonuniform DIF, power, robustness.

1. Introduction.

The topic of differential item functioning (DIF) has become a central concern in the measurement literature because tests are widely used in selection, promotion, certification and licensure decisions on several areas, such as education or industry. DIF is said to exist when examinees with the same ability level, but from different groups, have different probabilities of success on a given item. Uniform and nonuniform DIF have been defined by Mellenbergh (1982). Uniform DIF exists when the probability of answering the item correctly is greater for one group than the other uniformly over all levels of ability: There is no interaction between ability level and group membership. Nonuniform DIF exists when the probability of answering the item correctly is not greater across all levels of ability for any group: There is interaction between ability level and group membership.

A variety of statistical procedures for detecting DIF have been developed (Camilli & Shepard, 1994; Millsap & Everson, 1993; Fidalgo, 1996a; Potenza & Dorans, 1995). The most popular method for DIF detection is the Mantel-Haenszel (MH) procedure proposed by Holland and Thayer (1988). This procedure is particularly attractive because can be used with fewer examinees, is easy to program, relatively inexpensive in terms of computer time, and has an associated statistical test of significance. However, there is one caveat: the MH procedure is not sensitive to nonuniform DIF (Fidalgo, 1996b; Fidalgo & Mellenbergh 1995; Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990). The items that the MH procedure was most likely to miss were symmetric nonuniform DIF items of medium difficulty. In the item response theory (IRT) symmetric nonuniform DIF is defined for a difference in the item discrimination (a) parameters between groups. Therefore, the item response functions (IRFs) for the two groups intersec at the common difficulty parameter (b). In this situation, for items of moderate difficulty (b values about 0) the ICCs cross close the middle of the ability range, so the differences between both groups essentially canceled each other out, and the MH is unable to detect an interaction of this type. But for items of low or high difficulty, the IRFs cross either at low or the high end of the ability range. For this reason, the difference between the two groups do not cancel each other out, and the DIF can be detected using the MH procedure. The asymmetric nonuniform DIF is defined as a difference between groups in both a and b parameters. Recently a modification of the MH procedure was designed to detect nonuniform DIF (Mazor, Clauser & Hambleton, 1994). The modification consists of calculating the MH statistics separately for low-performing group (examinees with a total score $=<$ mean of the test score distribution in the total sample) and high-performing group (examinees with a total score $>$ mean of the test score distribution in the entire sample). However, this modification has not been investigated extensively. Another DIF assessment techniques are the loglinear and logit models (Fidalgo, 1996b; Fidalgo & Mellenbergh, 1995; Fidalgo, Mellenbergh & Muñiz, 1998; Fidalgo & Paz, 1995; Kok, Mellenbergh & Van der Flier, 1985; Van del Flier, Mellenbergh, Adèr & Wijn, 1984). An advantage of logit models is that they allow to test the hypothesis of no interaction between the ability variable and the group variable for detecting nonuniform DIF. Previous research has shown that the MH procedure improves the power over the logit models for detecting uniform DIF (Fidalgo, Mellenbergh & Muñiz, 1998), but that the logit models improve the power over the MH procedure for detecting nonuniform DIF (Fidalgo & Mellenbergh, 1995).

The present study continues the examination of the problem of detecting nonuniform DIF. Using simulated data, three methods for DIF detection are compared with respect to their power and robustness: the standard MH procedure, the modification (MHNU) of the MH procedure, and the iterative logit method.

2.- Method.

2.1.- Data generation.

To create conditions that were representatives of those found in practice, the item parameters used in the simulation are realistic: They are from a study on 70 items of the 1985 administration of the Graduate Management Admission Test (Kingston, Leary & Wightman, 1988). Clauser, Mazor and Hambleton (1994) fitted the three-parameter logistic model to these data (the c-parameter was fixed at a value of .2). These parameters are reported in Table 1.

	Item parameters			Item parameters			Item parameters	
Item number	a	b	Item number	a	b	Item number	a	b
1	.44	.64	26	.06	.44	51	.51	.55
2	-.42	.75	27	-2.41	.27	52	.80	.53
3	1.39	1.04	28	3.47	.67	53	.16	.91
4	-1.17	.47	29	1.43	1.10	54	-.12	.84
5	.28	.61	30	-1.23	.59	55	1.29	.10
6	-1.47	.32	31	1.40	.95	56	.07	.69
7	.37	.72	32	2.27	.70	57	-.47	.44
8	.97	.76	33	.26	.71	58	-.45	.53
9	-1.11	.15	34	2.26	1.30	59	1.61	.69
10	-1.13	.28	35	.99	.22	60	-2.27	.24
11	.22	1.00	36	-1.73	.37	61	1.24	.61
12	-1.07	.30	37	.64	.49	62	1.41	.75
13	.03	.93	38	-1.12	.57	63	.09	.85
14	-.18	.83	39	-.91	.43	64	.59	.96
15	-1.61	.54	40	.33	1.05	65	.75	.59
16	-.91	.40	41	-2.09	.34	66	-.78	.63
17	.12	.34	42	.55	1.27	67	.85	1.16
18	-1.20	.38	43	-.19	.33	68	-1.70	.43
19	.94	.73	44	1.74	1.21	69	-.23	.43
20	1.22	.42	45	.23	.40	70	.24	.79
21	-2.25	.51	46	-1.05	.58			
22	-1.30	.61	47	.73	1.30			
23	1.23	.82	48	.19	.33			
24	-2.08	.46	49	1.15	.39			
25	.88	.47	50	.51	.78			

Table 1. Item Parameters Used to Generate the Data Sets. All cs Were Set to Be .2

Simulated datasets were generated from a three-parameter logistic model (3PLM). The ability of the reference and focal groups was normally distributed with mean 0 and standard deviation 1 N(0, 1). With known ability for each examinee and item parameter values, the subject's probability of giving a correct response under the 3PLM is:

$$P_i(\theta) = c_i + \left[\frac{\left(1 - c_i\right)}{1 + \exp^{Da_i\left(\theta - b_i\right)}} \right]$$

Where, D = 1.7. The item score, 0 or 1, was generated by selecting a random number from a uniform distribution (0, 1). If the random number was less than or equal to Pi(θ), the examinee received a score of 1, otherwise the examinee received a score of 0.

2.2.- Design.

Two factors were manipulated: sample size, and DIF size (.25 and .50). To study the effects of these factors on Type I error rate and power of DIF detection methods, 4 conditions were simulated. These conditions were obtained by crossing two combinations of sample size (200 and 1,000 examinees per group) with two levels of DIF-size (area values of .25 and .50). For each condition 50 data sets were generated, and each of the data sets was analyzed using the standard MH, the new variation, and the iterative logit method. The set of item parameters for the nonDIF items were the same in both, reference and focal groups. To generate tests with about 11% of items with DIF, in addition to 70 nonDIF items showed in Table 1, nine items were added (items 71 through 79).The ICC parameters for these items in the reference group were obtained by crossing three levels of a parameter (.55, .65, and .75), and three levels of b parameter (-1.0, 0.0 and 1.0). Nonuniform DIF was generated varying the a parameters of the focal group. The magnitude of DIF was quantified in terms of the unsigned area between the generating IRFs, using Raju's (1988) formula no. 8. The parameters for DIF items in the reference and focal groups are reported in Table 2.

		Reference group		Focal group	
DIF size	Item no.	b	a	b	a
0.25	71	-1.0	.25	-1.0	.2765
	72	-1.0	.50	-1.0	.6185
	73	-1.0	.75	-1.0	1.0526
	74	0.0	.25	0.0	.2765
	75	0.0	.50	0.0	.6185
	76	0.0	.75	0.0	1.0526
	77	1.0	.25	1.0	.2765
	78	1.0	.50	1.0	.6185
	79	1.0	.75	1.0	1.0526
0.50	71	-1.0	.25	-1.0	.3093
	72	-1.0	.50	-1.0	.8107
	73	-1.0	.75	-1.0	1.7640
	74	0.0	.25	0.0	.3093
	75	0.0	.50	0.0	.8107
	76	0.0	.75	0.0	1.7640
	77	1.0	.25	1.0	.3093
	78	1.0	.50	1.0	.8107
	79	1.0	.75	1.0	1.7640

Table 2. a- and b- Parameters in the Reference and Focal Group Used to Generate Items with Weak (.25) and Moderate (.50) Symmetric Nonuniform DIF

2.3.- Analyses.

The Mantel-Haenszel procedure. The basic data used by the MH method are in the form of m 2 x 2 contingency tables, being m the number of ability levels or score categories. There are two levels per group: the focal group (F), which is the focus of analysis, and the reference group (R), that serves as a basis for the comparison. The total test score is used as a measure of ability, and the focal and reference group are matching on it.

Figure 1 is the 2 x 2 contingency table for the k ability level. In the rows, R and F denote the reference and the focal group respectively. In the columns, 0 denotes that the studied item is answered incorrectly, or that item has been failed, and 1 denotes that it is answered correctly. The cell values Ak , Bk , Ck and Dk denote the number of examinees in each category. The marginal values NRk and NFk represent the number of examinees in the R and F group, respectively; and N1k and N0k represent the number of examinees who answer correctly and incorrectly the studied item, respectively. Finally, Nk is the total number of examinees at the k ability level.

Group	1	0	Total
R	A_k	B_k	N_{Rk}
F	C_k	D_k	N_{Fk}
	N_{1k}	N_{0k}	N_k

Figure 1. The 2 x 2 contingency table for the k ability level.

The MH measure of DIF calculated across all 2 x 2 contingency tables is the common odds ratio estimator ($\hat{\alpha}_{MH}$) , given by

$$\hat{\alpha}_{MH} = \frac{\sum_{k=1}^{m} A_k D_k / N_k}{\sum_{k=1}^{m} B_k C_K / N_k}$$

The range of $\hat{\alpha}_{MH}$ goes from 0 to ∞. A value of 1 represents the null hypotesis of no DIF. If $\hat{\alpha}_{MH}$ is greater than 1, the studied item is favoring the reference group, on the contrary if $\hat{\alpha}_{MH}$ is less than 1 the studied item is favoring the focal group. Holland and Thayer (1985) converted $\hat{\alpha}_{MH}$ into a difference in delta metric of item difficulty via:

$$MH \quad D-DIF = -2.35\ln\left(\hat{\alpha}_{MH}\right)$$

where ln ($\hat{\alpha}_{MH}$) denotes the natural logarithm of $\hat{\alpha}_{MH}$. A zero value of MH D-DIF denotes

no DIF, a negative value implies that the item favors reference group, and a positive value implies that item favors focal group.

The MH statistics, with a continuity correction, to test the null hypotesis of $\hat{\alpha}_{MH} = 1$ is given by

$$\chi^2_{MH} = \frac{\left(\left| \sum_{k=1}^{m} A_k - \sum_{k=1}^{m} E\left(A_k\right) \right| - 0.5 \right)^2}{\sum_{k=1}^{m} Var\left(A_k\right)}$$

Where $E(A_k)$ is the expectation of A_k and $Var(A_k)$ is the variance of A_k given by:

$$E\left(A_k\right) = \left(N_{Rk}\, N_{1k}\right) / \, N_k$$

and

$$Var\left(A_k\right) = \frac{N_{Rk}\, N_{Fk}\, N_k}{N^2\left(N_k - 1\right)}$$

The MH chi-square statistic (χ^2_{MH}) follows a chi-square distribution with one degree of freedom. If χ^2_{MH} exceeds the table value of the chi-square distribution at a specified level of significance, the studied item exhibits DIF.

The MH procedures were run for each dataset using a modification of the MHDIF computer program (Fidalgo, 1994). In the first analysis, the program computes the MH statistics in the standard way, this is, in the total sample. In the second analysis, it uses the modification of the MH procedure proposed by Mazor, Clauser and Hambleton (1994).

The iterative logit method. Logit models for DIF detection, as they were formulated by Mellenbergh (1982), were used for analizing the following three-way contingency table: Score Categories by Item Score by Group. Therefore, the data for each item are summarized in a Score x 2 x 2 contingency table (supposing item responses scored as correct or incorrect, and two groups). The observed frequency in the ith score category (i = 1, 2, . . ., s), jth group (j = 1, . . ., g; g = 2), and kth response category (k = 1 for a correct response, and k = 2 for a incorrect response) is denoted by f_{ijk} . The expected frequency in the ith score category (i = 1, 2, . . ., s), jth group (j = 1, . . ., g; g = 2), and kth response category (k = 1 for a correct

response, and k = 2 for a incorrect response) is denoted by F_{ijk}. The logit is defined as the natural logarithm of the ratio of correct and incorrect responses. The saturated logit model may be formulated as (Fienberg, 1980, chap. 6)

$$\ln(F_{ij1}/F_{ij2}) = C + S_i + G_j + (SG)_{ij}$$

with the following constraints:

$$\sum_{i=1}^{s} S_i = 0$$

$$\sum_{j=1}^{g} G_i = 0$$

$$\sum_{i=1}^{s} SG_{ij} = \sum_{j=1}^{g} SG_{ij} = 0$$

where ln denotes the natural logarithm, C is an item constant, S_i the main Score Category effect, G_j the main Group effect, and $(SG)_{ij}$ the Score Category x Group interaction effect parameter. The model of Equation 7 fits perfectly the data. Other two logit models are of interest (Mellenbergh, 1982). When the following model fits the data (with first constraint applying)

$$\ln(F_{ij1}/F_{ij2}) = C + S_i$$

the item does not display DIF. If the valid model is equal to [with first and second constraints applying]

$$\ln(F_{ij1}/F_{ij2}) = C + S_i + G_j$$

the item showed uniform DIF. When the $(SG)_{ij}$ term cannot be dropped from the model of Equation 7, this is, the interaction between test score and group is necesary so that model fits the data, the item showed nonuniform DIF. The parameters of the models and their asymptotic standars errors are estimated from a sample using the method of maximum likelihood. The fit of the model is assessed by computing the expected frequencies given the model and the likelihood ratio statistic, which are symptotically chi-square distributed. For the unbiased item model the expected frecuencies are computed by

$$F_{ijk} = \left(\sum_{j=1}^{2} f_{ijk} \right) \left(\sum_{k=1}^{2} f_{ijk} \right) \Bigg/ \left(\sum_{j=1}^{2} \sum_{k=1}^{2} f_{ijk} \right)$$

The likekihhod ratio statistic (G^2) is

$$G^2 = 2\sum_{i=1}^{s}\sum_{j=1}^{2}\sum_{k=1}^{2} f_{ijk} \ln\left(\frac{f_{ijk}}{F_{ijk}}\right)$$

which is asymptotically chi-square distributed with s(g - 1) degrees of freedom.

Van der Flier, Mellenberg, Adèr, and Wijn (1984) applyed the logit model to an iterative way. The iterative logit method, as it was described by Kok, Mellenbergh and Van der Flier (1985), is the following: The test scores (minus the scores on the item under investigation) are divided into a predetermined number of categories. In the firststep of the iterative procedure the logit model is fitted to all the items of the test. The item with the hihgest (significant) G^2 value are identified. In the second step this item is eliminated from the test, and subject groups are reformed using the total test scores (minus the scores on the item under investigation) in the reduced (n - 1)-item test. The logit model is again fitted to the data for all n items, and the two items with the highest (significant) G^2 values are identified. In the next step these two items are eliminated, subject groups are reformed using scores on the test (minus the scores on the item under investigation) of (n - 2) items, and the logit refit to all n items. The procedure stops when it has iterated a predetermined number of times or when all the items in the reduced test have nonsignificant G2 values.

3.- Results.

3.1.- Power.

		Detection method		
Sample size	DIF size	MH	MHNU	LOGIT
N = 200	.25	.09	.11	.05
	.50	.15	.22	.12
N = 1,000	.25	.15	.28	.17
	.50	.41	.64	.56

Table 3. Estimated Power of the DIF Detection Methods per Sample size and DIF size

The proportion of correctly identified DIF items in 50 data set was used as a power estimate. These estimations are reported in Table 3. The results showed that the power of the MHNU procedure was higher than the power of the standard MH procedure and the iterative logit method at each of the factor levels. The lowest increase in the detection rates for MHNU over the standard MH procedure (MH) was a difference of about 3% when N = 200 and the size of DIF is equal to .25. The greatest increase was a difference of about 22% when N = 1,000 and the DIF size is equal to .50. The lowest increase for MHNU over the iterative logit method was a difference of 6.44% when N = 200 and the DIF size is equal to .25; and the greatest increase was a difference of approximately 11% when N = 1,000 and the DIF size is

equal to .25. On the other hand, only when there were 1,000 examinees per group the iterative logit method were more powerful than the standard MH procedure, with an increase of about 15% in the detection rates when the DIF size is equal to .50, and 2% when the DIF size is equal to .25.

3.2.- Robustness.

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 Estimated Type I Error Rates Over 50 Replications Under All Conditions $\alpha = .05$ if the Type I error rate is between .025 and .075.

		Detection method		
Sample size	DIF size	MH	MHNU	LOGIT
N = 200	.25	.033	.059	.031
	.50	.035	.059	.027
N = 1,000	.25	.038	.085*	.034
	.50	.049	.085*	.032

Note: An asterisk indicates that the test does not fulfil Bradley's (1978) liberal robustness criterion.

Table 4. Estimated Type I Error Rates Over 50 Replications Under All Conditions

The estimated Type I error rates at nominal level $\alpha = .05$ are reported in Table 4. The table shows that for N = 200 each of the three methods for DIF detection fulfils Bradley's liberal criterion. However, MHNU showed inflated Type I error rates (.085) when N = 1,000.

4.- Discussion.

Nonuniform DIF results from an interaction between ability and group membership, as can be seen in the logit model that describe the nonuniform DIF. Since the MH method uses a signed statistic, positive differences in one part of the ability distribution can offset negatives ones in another, in consequence this procedure is unable to identify the symmetric nonuniform DIF. For this cause, Mazor, Clauser and Hambleton (1994) proposed splitting the sample into high and low performing samples, and later to compute the MH statistic on each group. Using simulated data Mazor, Clauser and Hambleton (1994) showed that the modification improved detection rates of nonuniform DIF over the standard MH procedure, without increasing the Type I error rate. Nevertheless, their efficacy over other methods for detecting nonuniform DIF was not tested. The results of this study suggest that the modification of the MH procedure is more powerful than the standard MH procedure and the iterative logit method in detecting symmetric nonuniform DIF under all the factors manipulated here. As expected, sample size and DIF size had a large influence on power. In particular, the combination of a N = 1,000 sample size and a .50 DIF size showed the highest estimated power (a value of .41, .64, .56 for MH-, MHNU- and logit-tecniques, respectively). However, the estimated power at each of the factor levels and detection method is considerable low ranged from .09 to .64. These results are

in agreement with those obtained by other researchs that showed that poorly discriminating items may be difficult to detect (Clauser, Mazor & Hambleton, 1991; Fidalgo, 1996b; Fidalgo, Mellenbergh & Muñiz, 1999; Hambleton, Clauser, Mazor & Jones, 1993; Mazor, Clauser & Hambleton, 1994; Narayanan & Swaminathan, 1994, 1996), and for the 33% of DIF items the a-parameter value was set at .25. The investigation of the Type I error rates indicated that both the standard MH procedure and the iterative logit metod are robust. Whereas, the MHNU procedure showed a inflated Type I error rate (.085 at nominal level $\alpha = .05$) under the N = 1,000 sample size. For practicioners, this result suggests that considerable caution should be used in interpreting the results obtained from the MHNU procedure.

Our findings here are limited to the manipulated variables and items parameters used in this study, and several areas merit further investigation. Other procedures for nonuniform DIF detection are the logistic regresion (LR) procedure (Swaminathan & Rogers, 1990) and the crossing-SIBTEST (Li & Stout, 1996). The LR procedure can be conceptualized as being a logit model in wich the ability variable is treated as continuous (Fidalgo, 1996a, pp.409-411). Therefore, LR procedure is expected to improve on the iterative logit method for DIF detection. Recently, the results of a simulation study showed that the LR procedure is more powerful than the MHNU procedure in detecting nonuniform DIF (an overall increase of approximately 5%), but that it showed a substantial number of robustness violations (Ferreres, Fidalgo & Muñiz, 2000). On the other hand, Narayanan and Swaminathan (1996) compared the LR and crossing-SIBTEST procedures. Their results showed that overal there was high power and agreement between both crossing-SIBTEST and LR in detecting nonuniform DIF, but both procedures has inflated Type I error rate. In this way, it may be appropiate to compare the performance of the MHNU-, Crossing SIBTEST- and LR-procedures in detecting nonuniform DIF.

5.- 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.

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

Clauser, B., Mazor, K. M. & Hambleton, R. K. (1994). The effects of score group width on the Mantel-Haenszel procedure. *Journal of Educational Measurement, 31*, 67-78.

Ferreres, D., Fidalgo, A.M. & Muñiz, J. (2000). Detección del funcionamiento diferencial de los items no uniforme: Comparación de los métodos Mantel-Haenszel y regresión logística. *Psicothema, 12*, Supl. nº 2, 220-225.

Fidalgo, A. M. (1994). MHDIF: A computer program for detecting uniform and nonuniform differential item functioning with the Mantel-Haenszel procedure. *Applied Psychological Measurement, 18*, 300.

Fidalgo, A.M. (1996a). Funcionamiento diferencial de los items. In J. Muñiz (Ed.), *Psicometría* (pp. 371-455). Madrid, Spain: Universitas.

Fidalgo, A. M. (1996b). *Funcionamiento diferencial de los items: Procedimiento Mantel-Haenszel y modelos loglineales*. Unpublished doctoral dissertation, University of Oviedo.

Fidalgo, A. M. & Mellenbergh, G. J. (1995, April). *Evaluación del procedimiento Mantel_Haenszel frente al método logit iterativo en la detección del funcionamiento diferencial de los items uniforme y no uniforme*. Paper presented at the IV Symposium de Metodología de las Ciencias del Comportamiento, LaManga del Mar Menor, Spain.

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. *Psicothema, 10*, 219-228.

Fidalgo, A. M., Mellenbergh, G.J. & Muñiz, J. (1999). Aplicación en una etapa, dos etapas e iterativamente de los estadísticos Mantel-Haenszel. *Psicológica, 20*, 227-242.

Fidalgo, A.M. & Paz, M.D. (1995). Modelos lineales logarítmicos y funcionamiento diferencial de los items. *Anuario de Psicología, 64*, 57-66.

Fienberg, S. E. (1980). *The analysis of cross-classified categorical data* (2nd ed.). Cambridge, MA: MIT Press.

Hambleton, R. K., Clauser, B. E., Mazor, K. M. & Jones, R. W. (1993). Advances in the detection of differencially functioning test items. *European Journal of Psychological Assessment, 9*, 1-18.

Holland, W. P. & Thayer, D. T. (1985). *An alternative definition of ETS delta scale of item difficulty (Research Report No. 85-43)*. Princenton, 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: LEA.

Kingston, N., Leary, L. & Wightman, L. (1988). *An exploratory study of the applicability of item response theory methods to the Graduate Management Admissions Test (GMAC Occasional Papers)*. Princenton, NJ: Graduate Management Admissions Council.

Kok, F. G., Mellenbergh, G. J. & Van Der Flier, H. (1985). Detecting experimentally induced item bias using the iterative logit method. *Journal of Educational Measurement, 22*, 295-303.

Li, H. & Stout, W. (1996). A new procedure for detecting crossing DIF. *Psychometrika, 61*, 647-677.

Mazor, K. M., Clauser, B. E. & Hambleton, R. K. (1994). Identification of nonuniform differential item functioning using a variation of the Mantel_Haenszel procedure. *Educational and Psychological Measurement, 54*, 284-291.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7*, 105-107.

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

Narayanan, P. y 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. y Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement, 20*, 257-274.

Potenza, M. T. & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement, 19*, 23-37.

Raju, N. S. (1988). The area between two item characteristics curves. *Psychometrika, 53*, 495-502.

Rogers, H. J. & Swaminathan, H. (1993). A comparison of logistic regression and Mantel_Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*, 105-116.

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

Van Der Flier, H., Mellenbergh, G. J., Adèr, H. J. & Wijn, M. (1984). An iterative item bias detection method. *Journal of Educational Measurement, 21*, 131-145.