A Comparison of Logistic Regression and Mantel-Haenszel Procedures for Detecting Differential Item Functioning

H. Jane Rogers, Teachers College, Columbia University

Hariharan Swaminathan, University of Massachusetts

The Mantel-Haenszel (MH) procedure is sensitive to only one type of differential item functioning (DIF). It is not designed to detect DIF that has a nonuniform effect across trait levels. By generalizing the model underlying the MH procedure, a more general DIF detection procedure has been developed (Swaminathan & Rogers, 1990). This study compared the performance of this procedure—the logistic regression (LR) procedure—to that of the MH procedure in the detection of uniform and nonuniform DIF in a simulation study which examined the distributional properties of the LR and MH test statistics and the relative power of the two procedures. For both the LR and MH test statistics, the expected distributions were obtained under nearly all conditions. The LR test statistic did not have the expected distribution for very difficult and highly discriminating items. The LR procedure was found to be more powerful than the MH procedure for detecting nonuniform DIF and as powerful in detecting uniform DIF. *Index terms: differential item functioning, logistic regression, Mantel-Haenszel statistic, nonuniform DIF, uniform DIF.*

The Mantel-Haenszel (MH) procedure is currently one of the most popular procedures for detecting differential item functioning (DIF). The primary reasons for its popularity include its computational simplicity, ease of implementation, and associated test of statistical significance. These advantages, however, are obtained at the cost of some generality. The MH procedure is designed to detect uniform DIF and may not be sensitive to nonuniform DIF.

Uniform DIF exists if there is no interaction between trait level and group membership when the probability of success on an item is expressed in the logit metric. That is, the log-odds ratio of success on the item is greater for one group uniformly over all trait levels. Conversely, nonuniform DIF exists when there is interaction between trait level and group membership. That is, the difference in the log-odds ratio for the two groups is not the same at all trait levels. Most of the currently available DIF procedures ignore the existence of nonuniform DIF. Given that the current understanding of the nature of item bias is at best incomplete, it seems somewhat premature to assume that nonuniform DIF does not occur and to focus only on methods for detecting uniform DIF.

As an alternative to the MH procedure and also to the more complex and costly item response theory (IRT) procedures (Hambleton & Swaminathan, 1985), Swaminathan & Rogers (1990) developed a logistic regression (LR) procedure for detecting DIF. This procedure is an extension of the MH procedure that is effective in detecting both uniform and nonuniform DIF, as demonstrated by Swaminathan and Rogers. This study examined the relative efficacy of the LR and MH procedures under a variety of conditions.

The LR Procedure

The LR model for DIF is given by (Swaminathan & Rogers, 1990)

$$P(u = 1) = \frac{e^z}{(1 + e^z)}, \quad (1)$$

where

$$z = \tau_0 + \tau_1\theta + \tau_2 g + \tau_3(\theta g) . \quad (2)$$

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 17, No. 2, June 1993, pp. 105-116

0146-6216/93/020105-12$1.85

In this model, θ is the *observed* trait level of the examinee (usually total test score), and g represents group membership, which is defined as follows:

$$g = \begin{cases} .5 \text{ if person is a member of group 1} \\ -.5 \text{ if person is a member of group 2} \end{cases}, \quad (3)$$

and θg is the product of the two independent variables, g and θ. The parameter τ_2 corresponds to the group difference in performance on the item, and τ_3 corresponds to the interaction between group and trait level. An item shows uniform DIF if $\tau_2 \neq 0$ and $\tau_3 = 0$, and nonuniform DIF if $\tau_3 \neq 0$ (whether or not $\tau_2 = 0$).

In the LR model, the hypothesis of interest is $\tau_2 = \tau_3 = 0$. Swaminathan & Rogers (1990) showed that the statistic for testing this hypothesis has an asymptotic χ^2 distribution with 2 degrees of freedom (*df*). When the value of the statistic exceeds $\chi^2_{\alpha;2}$, the hypothesis that there is no DIF is rejected. The item then can be marked for further study by content specialists.

Because the LR procedure contains parameters representing both uniform and nonuniform DIF, it may be less powerful than the MH procedure in detecting strictly uniform DIF. That is, the interaction term may adversely affect the power of the procedure when only uniform DIF is present because 1 *df* is unnecessarily lost. Conversely, the MH procedure is designed to detect uniform DIF and hence may not be effective in detecting nonuniform DIF. Two simulation studies were implemented to study the effects of these differences between the two procedures on their relative effectiveness.

Method

The first study examined the distributions of the test statistics of the LR and MH procedures. For the procedures to be effective in detecting DIF, they must satisfy the distributional assumptions on which they are based. The second study investigated the relative power of the two procedures to detect uniform and nonuniform DIF.

Study 1: The Distribution Study

To study the distributions of the test statistics of the LR and MH procedures, empirical sampling distributions were constructed under several conditions. Because the distribution of the LR test statistic was of greatest interest, factors that might affect the distribution of the LR test statistic were identified and manipulated. Two factors were selected: sample size and the degree of model-data fit. Only when the LR model fits the data will the asymptotic results be valid; even then, a sufficient sample size is required to guarantee the asymptotic results (as for the MH statistic). Note that because of the different derivation of the MH test statistic, model-data fit may not have the same effect on the distribution of the MH statistic as it does on the LR statistic.

To study the effects of these factors, four conditions were simulated. Two levels of model-data fit ("good" fit and "poor" fit) were crossed with two levels of sample size (250 per group and 500 per group). Test data for which the LR model provided "good" fit were generated using the two-parameter logistic IRT model (2PLM). Although data generated in this way reflected unknown θs for which test scores may be poor estimates, it was expected that the degree of model-data fit would be reasonably good. Test data for which the LR model provided "poor" fit were obtained by simulating data based on the three-parameter logistic model (3PLM): Misfit would arise because the LR model specifies a lower asymptote of 0. In generating data under the 3PLM, all c values were set at .2.

In generating response patterns, trait values for the desired number of examinees were drawn from a standard normal distribution, and item parameters were chosen for a 40-item test (which is approximately average length for subtests of standardized achievement tests) from a bank of parameter estimates obtained from the analysis of real test data. Item parameters were selected to produce an approximately normal distribution of test scores. The item and θ parameters were substituted into the appropriate IRT model to obtain a probability of correct response on each item for each examinee. These probabilities were converted to item responses by comparing each

probability with a random number from a uniform distribution on the interval [0, 1]. The item was scored correct if the probability exceeded the random number and 0 otherwise.

Response vectors were generated for two groups, reflecting reference and focal groups, using the same item parameters in each group. Hence, all items were unbiased. For each combination of sample size and model-data fit, 100 replications of the data were performed.

Five of the 40 items were selected in order to study the distributions of the test statistics over the 100 replications. The five items were chosen to vary in level of difficulty (b) and discrimination (a) because these characteristics can be expected to affect the estimation of parameters and hence the distribution of the test statistic of the LR model. The levels of b and a and the item parameters for the five items were:

1. low b, low a ($b = -1.5$, $a = .6$);
2. moderate b, moderate a ($b = 0$, $a = 1$);
3. high b, high a ($b = 1.5$, $a = 1.6$);
4. high b, low a ($b = 1.5$, $a = .6$);
5. low b, high a ($b = -1.5$, $a = 1.6$).

For each of these five items, the LR and MH test statistics were calculated, and empirical sampling distributions were constructed. The Kolmogorov-Smirnov test was performed to determine if the test statistics had the expected distributions.

Study 2: The Power Study

Factors that may affect the power of the LR and MH procedures were identified and manipulated. The factors selected were model-data fit, sample size, test length, the shape of the test score distribution, percent of items in which DIF occurred, type of item, and size of the DIF.

Model-data fit and sample size may affect the power of the LR and MH procedures because of their possible effects on the distributions of the test statistics, as discussed above. Test length affects the accuracy of total score as a measure of trait level: The longer the test, the more reliable the total score. Because total score is used as the predictor in the LR model and as the criterion for grouping examinees in the MH procedure, a more reliable score (longer test length) may result in improved estimates of the parameters for both procedures.

The shape of the test score distribution may affect the LR procedure because of its effect on the fit of the regression curve. As with any regression procedure, the curve will be best fitted when the predictor is distributed over its fullest possible range. When the test score distribution is skewed, there will be few predictor values at one extreme, possibly resulting in poor estimation, and hence reduced power for the LR procedure. For the MH procedure, the effect of a skewed test score distribution may be small. Cells in which there are no examinees will simply be skipped in the computation of the MH statistic.

The percent of items with DIF may affect the power of both procedures because of the effect on test score. The greater the percent of items with DIF in the test, the poorer the test score will be as a measure of trait level, and the poorer it will be as a predictor in the LR procedure and as a blocking factor in the MH procedure.

The type of item in which DIF occurs may also affect the power of the LR and MH procedures to detect DIF. As explained in Study 1, the level of difficulty and discrimination of the item may affect parameter estimation, and hence DIF detection, under the LR procedure.

The effect of the size of DIF is clear: The greater the DIF, the easier it should be to detect. For this study, the size of DIF in an item was quantified in terms of the area between the generating item response functions (IRFs). Area was calculated using the formula given by Raju (1988). (Because all c values were the same for both groups, the formula was appropriate in all cases.)

To study the effects of these factors on the relative detection rates of the LR and MH procedures, 32 conditions were simulated. These conditions were obtained by crossing two levels of model-data fit (good and poor fit, simulated as in Study 1 using the 2PLM and 3PLM), two levels of sample size (250 per group and 500 per group),

two levels of test length (40 items and 80 items), two levels of the shape of the test score distribution (normal and negatively skewed), and two levels of percent of items with DIF (15% including the item of interest and 0% other than in the item of interest). Within each condition, both uniform and nonuniform DIF were simulated. For each type of DIF in each condition, four sizes of DIF were studied, corresponding to area values of .2, .4, .6, or .8.

In simulating uniform DIF, the *a* parameters for the two groups were kept the same but the *b* parameters for the two groups were different. 16 items showing uniform DIF were obtained by varying the level of the common *a* parameter (low or high), the level of the *b* parameters for the two groups (both low, both moderate, both high), and the size of the DIF (an area value of .2, .4, .6, or .8). The levels of the *b* and *a* parameters were manipulated because of the possible effects of these item characteristics on the estimation of parameters under the LR procedure, as explained earlier. The levels of *b* and *a* were not completely crossed with each other, because for some combinations it was difficult to obtain reasonable item parameter values for the two groups that would give the desired degree of DIF. Four types of item were studied: (1) low *b*, high *a*; (2) moderate *b*, low *a*; (3) moderate *b*, high *a*; and (4) high *b*, high *a*. For each type of item, items showing uniform DIF corresponding to area values of .2, .4, .6, and .8 were generated.

In simulating nonuniform DIF, the *b* parameters for the two groups were the same, but the *a* parameters for the two groups were different. 15 items showing nonuniform DIF were obtained by varying the level of the common *b* parameter (low, moderate, or high), the level of the *a* parameters for the two groups (both low or both high), and the size of the DIF (an area value of .2, .4, .6, or .8). As described above, the levels of *b* and *a* were not completely crossed because it was not always possible to find reasonable item parameter values that would yield the desired condition. Four types of items were studied: (1) low *b*, low *a*; (2) moderate *b*, low *a*; (3) moderate *b*, high *a*; and (4) high *b*, low *a*. For each type of item, items showing nonuniform DIF corresponding to area values of .2, .4, .6, and .8 were generated. For the case of a moderate *b*, high *a* item, it was not possible to find reasonable item parameter values that would give an area value of .8; hence, this combination was omitted. In addition to the 15 items with nonuniform DIF in these four categories, a fifth category of items was constructed. For items in the fifth category, both the *b* parameters and the *a* parameters for the two groups were different. This category of DIF items will be referred to as "mixed" nonuniform.

35 items with DIF were constructed. Tables 1 and 2 show the item parameter values used to generate items with DIF. Unbiased items were generated using item parameter values taken from real data and selected to produce (with normally distributed trait levels for both groups) either a normal or skewed test score distribution. To generate tests with 15% DIF, five items (one less than the six DIF items needed for a test length of 40) or 11 items (one less than the 12 DIF items needed for a test length of 80) were selected from the set of DIF items. These items were kept the same in all analyses and were included in the test solely to provide the desired degree of test score contamination. DIF statistics were not calculated for these items.

Each of the 35 DIF items to be studied was added separately into the test (to make up the 15% DIF), its DIF statistics were calculated, and it was removed from the test to be replaced by another of the items showing DIF. This procedure was used in order to provide the same context for studying each DIF item. Similarly, for the condition in which there was no DIF other than in the item of interest, each DIF item was separately added to the test. Each condition was replicated 20 times, and the percentages of items showing uniform and nonuniform DIF that were detected by each procedure were compared.

Table 1
Item Parameters Used to Generate
Two-Parameter Items with DIF

Type of DIF, Type of Item, and Item	Size of DIF	Group 1		Group 2	
		a_1	b_1	a_2	b_2
Uniform DIF					
Low *b*, High *a*					
1	.2	1.20	−1.61	1.20	−1.39
2	.4		−1.71		−1.29
3	.6		−1.81		−1.19
4	.8		−1.91		−1.09
Moderate *b*, Low *a*					
9	.2	.60	−.11	.60	.11
10	.4		−.21		.21
11	.6		−.31		.31
12	.8		−.41		.41
Moderate *b*, High *a*					
5	.2	1.20	−.11	1.20	.11
6	.4		−.21		.21
7	.6		−.31		.31
8	.8		−.41		.41
High *b*, High *a*					
13	.2	1.20	1.39	1.20	1.61
14	.4		1.29		1.71
15	.6		1.19		1.81
16	.8		1.09		1.91
Nonuniform DIF					
Low *b*, Low *a*					
17	.2	.70	−1.50	.85	−1.50
18	.4	.60		.87	
19	.6	.50		.81	
20	.8	.45		.82	
Moderate *b*, Low *a*					
21	.2	.70	0.00	.85	0.00
22	.4	.60		.87	
23	.6	.50		.81	
24	.8	.45		.82	
Moderate *b*, High *a*					
25	.2	1.10	0.00	1.70	0.00
26	.4	.90		1.70	
27	.6	.70		1.50	
High *b*, Low *a*					
28	.2	.70	1.50	.85	1.50
29	.4	.60		.87	
30	.6	.50		.81	
31	.8	.45		.82	
Mixed, Moderate *b*, High *a*					
32	.2	1.20	−.10	1.40	.10
33	.4	1.20	−.20	1.80	.20
34	.6	1.10	−.30	1.70	.30
35	.8	1.00	−.40	1.60	.40

Table 2
Item Parameters Used to Generate Three-Parameter Items With DIF (c Values For Both Groups Were Fixed at .2)

Type of DIF, Type of Item, and Item	Size of DIF	Group 1		Group 2	
		a_1	b_1	a_2	b_2
Uniform DIF					
Low b, High a					
1	.2	1.20	−1.63	1.20	−1.37
2	.4		−1.77		−1.24
3	.6		−1.89		−1.11
4	.8		−2.02		−.99
Moderate b, Low a					
9	.2	.60	−.13	.60	.13
10	.4		−.27		.27
11	.6		−.39		.39
12	.8		−.52		.52
Moderate b, High a					
5	.2	1.20	−.13	1.20	.13
6	.4		−.27		.27
7	.6		−.39		.39
8	.8		−.52		.52
High b, High a					
13	.2	1.20	1.37	1.20	1.63
14	.4		1.24		1.77
15	.6		1.11		1.89
16	.8		.99		2.02
Nonuniform DIF					
Low b, Low a					
17	.2	.60	−1.50	.74	−1.50
18	.4	.50		.74	
19	.6	.45		.79	
20	.8	.42		.89	
Moderate b, Low a					
21	.2	.60	0.00	.74	0.00
22	.4	.50		.74	
23	.6	.45		.79	
24	.8	.42		.89	
Moderate b, High a					
25	.2	1.10	0.00	1.70	0.00
26	.4	.80		1.65	
27	.6	.60		1.40	
High b, Low a					
28	.2	.60	1.50	.74	1.50
29	.4	.50		.74	
30	.6	.45		.79	
31	.8	.42		.89	
Mixed, Moderate b, High a					
32	.2	1.00	−.10	1.35	.10
33	.4	1.00	−.25	1.50	.25
34	.6	.80	−.35	1.50	.35
35	.8	.60	−.40	1.35	.40

Results

Distribution Study

The results of the distribution study are presented in Tables 3 and 4. The LR test statistic should be distributed as a χ^2 with 2 *df*; the MH statistic should be distributed as a χ^2 with 1 *df*. The mean and standard deviation of the LR statistic should therefore both be 2, and the corresponding values for the MH statistic should be 1.

Table 3 shows that the MH procedure produced a slightly larger number of cases in which the test statistic did not appear to have the expected distribution. These cases did not correspond to the same items across the four conditions, and were not the same as those that were problematic for the LR statistic. The model used to generate the data did not appear to have an effect on the distribution of the MH statistic. There may have been a slight effect due to sample size. For 250 per group, the distributional assumptions were less often met, and the Kolmogorov-Smirnov values were somewhat larger.

Despite the violations of the distributional assumptions, Table 4 shows that the numbers of false positives were in accordance with expectations. The notable exception for the LR statistic was in the case of Item 3 in the poor fit and 500 per group condition. The number of false positives was large under this condition, and there appeared to be a tendency for it to produce a relatively large number of false positives overall.

Table 3
Mean and Standard Deviation (SD) of LR and MH Indexes, and Kolmogorov-Smirnov (KS) Statistics and Their Estimated Probabilities (*p*) for Testing the Distributional Assumptions

Model-Data Fit, Group Size, and Item	LR				MH			
	Mean	SD	KS	*p*	Mean	SD	KS	*p*
Good Fit, 250 Per Group								
1	1.85	1.85	.08	.54	.64	1.07	.16	.01*
2	2.19	2.02	.10	.23	.83	1.15	.09	.38
3	1.75	1.48	.08	.54	.75	1.07	.12	.12
4	2.12	1.99	.08	.54	.67	1.05	.15	.02*
5	2.00	1.70	.08	.54	.85	1.35	.14	.04*
Good Fit, 500 Per Group								
1	1.85	1.58	.05	.96	.69	.90	.12	.10
2	2.26	2.26	.16	.01*	1.02	1.60	.06	.87
3	2.17	2.24	.06	.87	.99	1.59	.11	.18
4	2.04	2.10	.06	.87	.90	1.43	.10	.23
5	2.10	2.03	.07	.75	.99	1.68	.11	.18
Poor Fit, 250 Per Group								
1	1.88	1.66	.05	.96	.64	.96	.16	.01*
2	1.80	1.63	.06	.87	.75	1.13	.13	.08
3	2.56	2.02	.16	.01*	.89	1.25	.11	.18
4	2.12	1.84	.12	.12	.83	1.25	.15	.02*
5	1.91	1.78	.08	.54	.86	1.09	.05	.96
Poor Fit, 500 Per Group								
1	2.04	2.20	.07	.75	.73	.99	.10	.23
2	1.87	2.03	.08	.54	.82	1.46	.09	0.00*
3	2.97	2.71	.18	0.00*	.98	1.63	.09	.38
4	2.04	2.05	.05	.96	.79	1.17	.10	.23
5	1.86	1.83	.08	.54	1.01	1.43	.08	.54

*$p < .05$.

Table 4
Number of Unbiased Items Falsely Identified by the LR and MH Procedures Using 95th and 99th Percentile Cutoffs

Model-Data Fit, Group Size, and Item	LR		MH	
	95	99	95	99
Good Fit, 250 Per Group				
1	6	0	3	0
2	6	1	3	0
3	1	0	2	0
4	5	1	2	0
5	3	0	5	0
Good Fit, 500 Per Group				
1	2	0	0	0
2	3	1	5	1
3	8	1	4	2
4	3	3	5	1
5	4	1	6	2
Poor Fit, 250 Per Group				
1	4	0	2	0
2	3	0	3	0
3	8	0	4	0
4	3	1	4	0
5	4	1	2	0
Poor Fit, 500 Per Group				
1	7	1	2	0
2	6	2	5	2
3	17	5	5	3
4	6	2	3	1
5	4	1	5	2

Power Study

To facilitate analysis of the detection rates for the LR and MH procedures, the data were analyzed by ANOVA in which the dependent variable was the number of times out of the 20 replications that the item containing DIF was identified as biased by each procedure, and the independent variables were the factors manipulated in the study. Because of the large cell sizes, it was expected that many of the effects would be statistically significant without being meaningful. On the other hand, given the cell sizes, nonsignificant effects could be meaningfully interpreted; therefore, equal attention was given to these. Detection rates for uniform and nonuniform DIF were analyzed separately.

Uniform DIF. The ANOVA results are presented in Table 5, and the corresponding means are given in Table 6. Before interpreting the main effects, the significant interactions were examined. Examination of the means corresponding to the interactions showed that, for the most part, the interactions were ordinal. For the disordinal interactions, the reversal in direction of the mean differences was not large enough to cancel out the main effect. Therefore, in those cases in which there was not a significant main effect, it was not an artifact of interaction.

Table 5 shows that for both procedures test length and shape of the score distribution did not appear to affect detection rates. For the LR procedure, model-data fit did not affect the results; for the MH procedure, model-data fit produced a significant effect (see Table 5), but the mean difference was small (see Table 6). Conversely, for the MH procedure, Table 5 shows that the percent of DIF items did not affect the results, but for the LR procedure it did. On average, the detection rate rose from 70% to 76% for the uniform LR procedure when the percent of items with DIF dropped from 15% to none other than in the item under study (see Table 6).

Table 6 shows that sample size appeared to have a strong effect on the detection rates for both procedures. Detection rates increased by approximately 15% when sample size was increased from 250 to 500. The type of item also had a large effect for both procedures. The items with DIF that were most easily detected by both procedures were items of moderate difficulty and high discrimination. For these items, detection rates were as much as 15% greater than for other types of items.

Size of DIF produced the expected effect—detection rates for both procedures were low (approximately 30%) for .2 DIF, but were very high (95%) for .6 DIF. Overall, the MH procedure had slightly higher detection rates for uniform DIF.

Nonuniform DIF. Results for the detection of nonuniform DIF also are presented in Tables 5 and 6. The results were quite different from those for uniform DIF. Table 5 shows that both procedures appeared to be unaffected by the shape of the score distribution or the percent of items with

Table 5
Analysis of Variance of the Effects of All Factors on the Performance of the LR and MH Procedures in Detecting Uniform and Nonuniform DIF

	Uniform				Nonuniform			
	LR		MH		LR		MH	
Factor	*F*	*p*	*F*	*p*	*F*	*p*	*F*	*p*
Model-Data Fit (MDF)	.74	.39	63.76	0.00	272.96	0.00	76.00	0.00
Sample Size (*N*)	591.56	0.00	677.93	0.00	448.00	0.00	164.65	0.00
Test Length (TL)	2.06	.15	2.13	.15	40.29	0.00	6.00	.15
Score Distribution (ScD)	.25	.62	.94	.33	.14	.71	.54	.71
Percent DIF (%DIF)	143.56	0.00	4.63	.03	3.53	.06	5.53	.06
Type of Item (Type)	119.92	0.00	128.24	0.00	353.00	0.00	913.06	0.00
Size of DIF (DIFSize)	3,031.28	0.00	2,376.11	0.00	1,017.27	0.00	646.52	0.00
MDF × *N*	.25	.62	.08	.78	.31	.58	2.61	.11
MDF × TL	.02	.89	0.00	.99	.11	.74	2.31	.13
MDF × ScD	.87	.35	2.13	.15	.90	.34	.09	.77
MDF × %DIF	.87	.35	0.00	.99	.41	.52	.19	.66
MDF × Type	25.12	0.00	81.60	0.00	26.08	0.00	21.67	0.00
MDF × DIFSize	.75	.48	3.15	.04	12.24	0.00	1.02	.36
N × TL	.25	.62	2.33	.13	0.00	.98	0.00	.96
N × ScD	.02	.89	.12	.71	.01	.91	.35	.56
N × %DIF	1.01	.32	1.23	.27	.60	.44	1.06	.30
N × Type	25.31	0.00	19.48	0.00	11.59	0.00	34.04	0.00
N × DIFSize	64.24	0.00	56.54	0.00	5.51	0.00	1.73	.18
TL × ScD	1.16	.28	0.00	.95	1.37	.24	2.16	.14
TL × %DIF	3.75	.05	1.56	.21	.41	.52	.19	.66
TL × Type	4.25	.01	3.43	.02	2.50	.04	1.55	.19
TL × DIFSize	1.37	.26	4.15	.02	1.66	.19	3.34	.04
ScD × %DIF	.33	.57	.12	.73	.06	.81	0.00	.99
ScD × Type	.32	.81	.59	.62	.21	.93	.16	.96
ScD × DIFSize	.31	.73	.37	.69	.17	.85	.09	.92
%DIF × Type	9.61	0.00	3.31	.02	11.98	0.00	.32	.86
%DIF × DIFSize	23.42	0.00	.11	.90	.02	.98	.31	.74
Type × DIFSize	42.54	0.00	18.97	0.00	45.42	0.00	112.78	0.00

DIF. The MH procedure was, again, not sensitive to test length, but in this case, the LR procedure was sensitive (Table 5), although the effect was fairly weak (Table 6).

Table 6 shows that both procedures were affected to some extent by model-data fit. The detection rate for the LR procedure was 14% lower when model-data fit was poor than when model-data fit was good; the detection rate for the MH procedure decreased 7%. Sample size produced a strong effect—LR detection rates increased 19% as sample size increased, and MH detection rates increased 11%.

Other than the size of the DIF, the largest effect observed for both procedures was due to the type of item. Table 6 shows that for the LR procedure, the lowest detection rate occurred with items of moderate difficulty and low discrimination, and the highest detection rate occurred for items of moderate difficulty and high discrimination. The MH procedure was almost completely unable to detect strictly nonuniform DIF in items of moderate difficulty. For items of low difficulty, the MH detection rate was still approximately 15% lower than the LR detection rate; for items of high difficulty, the detection rates were nearly the same. When the DIF was mixed nonuniform (differences in *b* and *a* for the two groups), the MH procedure was able to detect it as well as the LR procedure.

Table 6
Mean Percent Detection Rates for Uniform and Nonuniform DIF Over 20 Replications for the LR and MH Procedures Under All Conditions

Factor and Level	Uniform		Nonuniform	
	LR	MH	LR	MH
Model-Data Fit				
Good	73	71	67	44
Poor	73	73	53	37
Sample Size				
250 Per Group	66	68	50	35
500 Per Group	80	83	69	46
Test Length				
40	73	76	57	39
80	73	75	62	41
Score Distribution				
Normal	73	75	60	40
Skewed	73	76	60	41
Percent DIF				
15%	70	75	59	41
0%	76	76	61	40
Type of Item				
Low *b*, Low *a*			69	56
Low *b*, High *a*	71	73		
Moderate *b*, Low *a*	67	73	37	5
Moderate *b*, High *a*	82	85	71	5
High *b*, Low *a*			47	46
High *b*, High *a*	73	70		
Mixed, Moderate *b*, High *a*			78	80
Size of DIF				
.2	27	33	24	12
.4	71	75	57	38
.6	95	95	75	49
.8	99	99	86	68

Discussion

Distribution Study

Overall, the distributional assumptions of both procedures appeared to be met to a satisfactory degree, at least for practical purposes. The one exception that occurred for the LR procedure was with an item that was highly difficult and highly discriminating (Item 3). This result may be due to the greater influence of the *c* parameter on difficult items. For very easy items, the *c* parameter has an effect only on the very lowest part of the trait scale, and hence the LR model will fit the data adequately over nearly all the range. For very difficult and discriminating items, the *c* parameter affects a much larger part of the trait scale, hence the misfit of the LR model will become more apparent. In practical terms, this problem may not be serious, at least for most achievement tests, in which there are few very difficult but highly discriminating items.

Power Study

The LR and MH procedures were almost equally effective in detecting uniform DIF. The slight advantage of the MH procedure in detecting uniform DIF is probably due to its greater power for this purpose, gained by conserving 1 *df*. The gain, however, is quite small. The considerably better overall performance of the LR

procedure relative to the MH procedure makes the loss of 1 *df* well worthwhile. For example, for nonuniform DIF greater than .2, over all items and all conditions with the poor data, the LR procedure detected 57% of items with DIF and the MH procedure detected 34%.

The most noteworthy result was that concerning the differential effect of the type of item on the detection rates for the two procedures. The reason for this difference may be found by considering the IRFs for the two groups on the different types of items. For strictly nonuniform DIF, the IRFs for the two groups intersect at the common difficulty. Hence, for items of moderate difficulty, the IRFs cross in the center of the trait range. This situation may be thought of as disordinal interaction between trait level and group membership. Because the MH procedure is primarily sensitive to the main effect of group membership, it is unable to detect an interaction of this type. For items of low or high difficulty, the IRFs cross either at the low end or the high end of the trait scale; this situation reflects ordinal interaction. Because the main effect of group membership can be detected in the presence of ordinal interaction, the MH procedure can detect nonuniform DIF in items of this type under some conditions.

An interesting finding was that the percent of items with DIF in the test did not affect the MH results. This may be due to the two-stage process used in the MH procedure. In the first stage, examinees are grouped according to total score based on all items, and the MH procedure is used to identify items showing DIF. In the second stage, items showing DIF (with the exception of the item being studied) are excluded from the calculation of total score used to group examinees. Then the MH analysis is repeated. Because the percent of items in the test with DIF did not affect the MH results, it suggests that this two-stage "purification" process is effective. Given this result, it may be worthwhile to use such a purification process with the LR procedure.

Conclusions

The LR model is more general than the model underlying the MH procedure and hence is sensitive to DIF of a more general nature. The MH procedure, although computationally simple, is sensitive only to DIF that is approximately constant across all trait levels. This study demonstrated that the LR procedure is as powerful as the MH procedure in detecting uniform DIF, and more powerful than the MH procedure in detecting nonuniform DIF.

The MH procedure, however, is quick and inexpensive to implement. Calculation of the MH statistic and accompanying odds-ratio requires only cell frequencies. Estimation of parameters in the LR model is necessarily iterative, and hence more expensive in terms of computer time. The LR procedure appears to be three to four times more expensive than the MH procedure; nevertheless, this is still very inexpensive compared to IRT procedures.

For tests that are very easy or very difficult, the MH procedure might provide satisfactory results in detecting nonuniform DIF. When the test is of moderate difficulty, as are most achievement tests, the MH procedure might fail to detect instances even of large DIF. Because by the very use of the term DIF, the concern is with differential item functioning, not merely differential item difficulty, it seems inconsistent to focus on only one type of DIF because it can be detected with little effort or expense. The results of this study clearly show that there is a feasible alternative.

References

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory.* Boston MA: Kluwer-Nijhoff.

Holland, P. W., & 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.

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

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

Author's Address

Send requests for reprints or further information to H. Jane Rogers, School of Education, Hills House South, University of Massachusetts, Amherst MA 01003, U.S.A.