Chapter 15
Assessing Differential Item Functioning in Multiple Grouping Variables with Factorial Logistic Regression

Kuan-Yu Jin, Hui-Fang Chen, and Wen-Chung Wang

Abstract Differential item functioning (DIF) can occur among multiple grouping variables (e.g., gender and ethnicity). For such cases, one can either examine DIF one grouping variable at a time or combine all the grouping variables into a single grouping variable in a test without a substantial meaning. These two approaches, analogous to one-way analysis of variance (ANOVA), are less efficient than an approach that considers all the grouping variables simultaneously and decomposes the DIF effect into main effects of individual grouping variables and their interactions, which is analogous to factorial ANOVA. In this study, the idea of factorial ANOVA was applied to the logistic regression method for the assessment of uniform and nonuniform DIF, and the performance of this approach was evaluated with simulations. The results indicated that the proposed factorial approach outperformed conventional approaches when there was interaction between grouping variables; the larger the DIF effect size, the higher the power of detection; the more DIF items in the anchored test, the worse the DIF assessment. Given the promising results, the factorial logistic regression method is recommended for the assessment of uniform and nonuniform DIF when there are multiple grouping variables.

Keywords Differential item functioning • Logistic regression • Uniform differential item functioning • Nonuniform differential item functioning

Many tests and inventories have been developed to measure latent traits in the human sciences and to compare inter-individual differences. A major concern

K.-Y. Jin • W.-C. Wang
Assessment Research Centre, Hong Kong Institute of Education, 10 Lo Ping Road, Tai Po, New Territories, Hong Kong SAR
e-mail: kyjin@ied.edu.hk; wcwang@ied.edu.hk

H.-F. Chen (✉)
Department of Applied Social Sciences, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong SAR
e-mail: hfchen@cityu.edu.hk

R.E. Millsap et al. (eds.), *Quantitative Psychology Research*, Springer Proceedings in Mathematics & Statistics 89, DOI 10.1007/978-3-319-07503-7_15

that arises under such group comparisons is whether or not test items reflect the same latent dimensions across all groups of examinees, termed measurement equivalence or measurement invariance (Candell and Hulin 1986; Drasgow 1987). A lack of measurement invariance leads to a problematic situation where examinees having the same underlying ability but belonging to different groups have different probabilities of success on an item. Thus, the test favors one or more groups of examinees but disadvantages others. Measures are not comparable across groups, and test fairness is threatened.

Assessment of differential item functioning (DIF) is a routine practice to investigate measurement invariance at the item level, especially for large-scale assessment programs such as the Program for International Student Assessment and the Trends in International Mathematics and Science Study. DIF refers to examinees with the same ability level from different groups having different probabilities of pass or endorsing an item. In the framework of item response theory (IRT), an item shows DIF if its response functions are not identical across groups. The psychometric properties differ across groups, and the differences in the measures across groups do not reflect true differences.

Most DIF studies focus on the difference between a reference group (e.g., majority) and a focal group (e.g., minority). Latent traits of the two groups of examinees are placed on the same metric based on an anchored test, and then the responses to a studied item are examined for DIF. Sometimes, more than two groups of examinees may be involved, such as in cross-cultural and cross-ethnic research (Iwata et al. 2002). In such cases, a group (e.g., white Americans) is selected to serve as the reference group, so the other focal groups can be compared against the reference group, one focal group at a time. This procedure is analogous to the independent-samples *t*-test. Just as the one-way ANOVA is statistically superior to multiple independent-samples *t*-tests, simultaneous DIF analysis across multiple groups has been found to be statistically more efficient than multiple two-group DIF analyses (Güler and Penfield 2009; Kim et al. 1995; Penfield 2001).

Specifically, Kim et al. (1995) developed the Q_j statistic using the vectors of item parameter estimates. If the vectors differ significantly across groups, then the item characteristic functions differ across groups, and the item is deemed to exhibit DIF. Being an IRT-based method, the Q_j statistic requires large sample sizes for stable item parameter estimation. To resolve this problem, Penfield (2001) proposed a non-IRT-based method: the generalized Mantel–Haenszel (MH) statistic (Somes 1986; Zwick et al. 1993). Simulation results confirmed that both methods yielded well-controlled Type I error rates and high power rates, but they differed in computation time and sample size requirements.

When DIF analysis is to be conducted on multiple grouping variables (factors), such as gender (two levels) and ethnicity (three levels), two approaches are often adopted: The first approach is to consecutively conduct DIF analysis, one grouping factor at a time. For example, one can conduct a gender DIF analysis, followed by an ethnicity DIF analysis. The second approach is to combine these two grouping factors into a pseudo-grouping factor with six levels and to implement the procedures proposed by Kim et al. (1995) or Penfield (2001). The first approach, analogous to conducting one-way ANOVA procedures consecutively, aims to evaluate whether

there is a gender DIF or an ethnicity DIF. The second approach, also analogous to one-way ANOVA, creates a pseudo-grouping factor that often lacks substantial meaning. Both approaches are less statistically efficient than factorial ANOVA, where all grouping factors are simultaneously considered and the "total" DIF effect is partitioned into main effects of individual grouping factors and their interaction effects, such as a main effect of gender, a main effect of ethnicity, and an interaction effect between gender and ethnicity.

Factorial DIF analysis procedures in the framework of Rasch models have been proposed and proven to be effective in DIF assessment (Wang 2000a, b) and outperform conventional consecutive DIF analyses when an interaction exists between grouping factors (Chen et al. 2012). Embedded in the framework of Rasch models, such factorial procedures are parametric and not applicable to the assessment of nonuniform DIF. In this study, we adopt the logic of factorial DIF analysis and apply it to a nonparametric approach—the logistic regression (LR) method (Swaminathan and Rogers 1990)—which is applicable to both uniform and nonuniform DIF.

The LR method is one of the most widely used nonparametric approaches in DIF assessment (Kim and Oshima 2013; Li et al. 2012). It is simple, easy to implement, and does not require a large sample size or a specific form of item response functions. It can be easily implemented in common computer packages such as SPSS, SAS, or Matlab, or free software such as R. The LR method works equally as well as the MH method in uniform DIF assessment, and outperforms the MH method in nonuniform DIF assessment (Narayanan and Swaminathan 1994, 1996; Swaminathan and Rogers 1990). Often, a raw test score is treated as a matching variable to place examinees from different groups on the same metric, so studied items can be assessed for uniform or nonuniform DIF. Compared to IRT-based DIF assessment methods, disadvantages of the LR method include inflated Type I error rates when different groups of examinees have very different mean ability levels (Güler and Penfield 2009; Narayanan and Swaminathan 1996) and its poor performance when the underlying IRT model is a multiparameter logistic model (Bolt and Gierl 2006; DeMars 2010).

Given the importance of factorial DIF analysis and the simplicity and popularity of the LR method in uniform and nonuniform DIF assessment, this study develops the factorial logistic regression (FLR) method to assess DIF effects when there are multiple grouping factors. Its performance in DIF assessment is evaluated and compared to other LR methods via two simulation studies. In the following sections, we introduce the key ideas of the FLR method, present the results of the simulation studies, draw conclusions, and give suggestions for future studies.

15.1 The FLR Method

Let T_n denote the raw test score for person n. Let X_n be an indicator of group membership for person n; for example, $X_n = 1$ if person n belongs to the reference group, and $X_n = -1$ if person n belongs to the focal group. Let P_n be the probability of

success on the studied item for person n. When the studied item is to be assessed for DIF, one can formulate the log-odds (or logit) of a correct answer over an incorrect answer as:

$$\log\left(\frac{P_n}{1-P_n}\right) = \tau_0 + \tau_1 T_n + \tau_2 X_n + \tau_3 X_n T_n, \tag{1}$$

where τ_0 - τ_3 are the regression coefficients for the studied item. If τ_2 or τ_3 is not zero, then the item is deemed to exhibit DIF. Normally, if τ_3 is not zero, then the item is deemed to exhibit nonuniform DIF; if τ_3 is zero but τ_2 is not, then the item is deemed to exhibit uniform DIF (Narayanan and Swaminathan 1994).

When there is one grouping factor and it has more than two groups ($g = 1, \ldots, G$), one can create a set of G - 1 dummy variables to represent the group membership: $\mathbf{X}_n{}' = (X_{n1}, \ldots, X_{n(G-1)})$. For example, if there are three groups, two dummy variables, X_1 and X_2, can be created. If examinee n is in group 1, then $X_{n1} = 1$, $X_{n2} = 0$; in group 2, $X_{n1} = 0$, $X_{n2} = 1$; in group 3, $X_{n1} =$ -1, $X_{n2} =$ -1. That is,

$$\mathbf{X}_n{}' = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{bmatrix}, \tag{2}$$

where the two columns stand for X_1 and X_2, and the three rows stand for the three groups. Equation (1) can then be extended as follows:

$$\log\left(\frac{P_n}{1-P_n}\right) = \tau_0 + \tau_1 T_n + \boldsymbol{\tau}_2{}'\mathbf{X}_n + \boldsymbol{\tau}_3{}'\mathbf{X}_n T_n, \tag{3}$$

where τ_0, τ_1, $\boldsymbol{\tau}_2$, and $\boldsymbol{\tau}_3$ are the regression coefficients for the studied item. For the three groups, Eq. (3) becomes

$$\log\left(\frac{P_n}{1-P_n}\right) = \tau_0 + \tau_1 T_n + \tau_{21} X_{n1} + \tau_{22} X_{n2} + \tau_{31} X_{n1} T_n + \tau_{32} X_{n2} T_n, \tag{4}$$

where $\boldsymbol{\tau}_2{}' = (\tau_{21},\ \tau_{22})$, $\boldsymbol{\tau}_3{}' = (\tau_{31},\ \tau_{32})$, and $\mathbf{X}_n{}' = (X_{n1}, X_{n2})$. If $\boldsymbol{\tau}_3$ is not a zero vector, then the item is deemed to exhibit nonuniform DIF; if $\boldsymbol{\tau}_3$ is a zero vector but $\boldsymbol{\tau}_2$ is not, then the item is deemed to exhibit uniform DIF.

The interpretation of $\boldsymbol{\tau}_2$ and $\boldsymbol{\tau}_3$ is analogous to that in standard logistic regression. Take the design matrix in Eq. (3) as an example. When there is no nonuniform DIF (i.e., $\boldsymbol{\tau}_3 = \mathbf{0}$), then Eq. (4) becomes

$$\text{Group 1 } (X_1 = 1,\ X_2 = 0): \log\left(\frac{P_n}{1-P_n}\right) = \tau_0 + \tau_1 T_n + \tau_{21}, \tag{5}$$

$$\text{Group 2 } (X_1 = 0,\ X_2 = 1): \log\left(\frac{P_n}{1-P_n}\right) = \tau_0 + \tau_1 T_n + \tau_{22}, \tag{6}$$

$$\text{Group 3 } (X_1 = -1,\ X_2 = -1): \log\left(\frac{P_n}{1-P_n}\right) = \tau_0 + \tau_1 T_n - \tau_{21} - \tau_{22}. \tag{7}$$

If $\boldsymbol{\tau}_2{}' = (\tau_{21}, \tau_{22}) = (0.4, -0.3)$, then for examinees with an equal ability level, the log-odds (logit) of group 1 examinees will be 0.8 higher than that of group 3 examinees, and the log-odds (logit) of group 2 examinees will be 0.6 lower than that of group 3 examinees.

Next, suppose there is more than one grouping factor. For illustrative simplicity, let there be two grouping factors, A (e.g., gender) and B (e.g., ethnicity), and let each factor have two levels (e.g., male and female; white and black), so that in total there are four groups of examinees (e.g., white male, white female, black male, and black female). Let X_1 be the dummy variable for factor A, and X_2 be the dummy variable for factor B. To account for the interactions between factors A and B, one additional dummy variable is needed: X_1X_2. Thus, a 4 by 3 matrix can be created:

$$\mathbf{X}_n{}' = \begin{bmatrix} 1 & 1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & -1 \\ -1 & -1 & 1 \end{bmatrix}, \tag{8}$$

where the three columns stand for X_1, X_2, and X_1X_2, and the four rows stand for the four groups. That is, $X_{n1} = 1$, $X_{n2} = 1$, $X_{n1}X_{n2} = 1$ if examinee n is in group 1 (white male); $X_{n1} = $ -1, $X_{n2} = 1$, $X_{n1}X_{n2} = $ -1 if in group 2 (white female); $X_{n1} = 1$, $X_{n2} = $ -1, $X_{n1}X_{n2} = $ -1 if in group 3 (black male); $X_{n1} = $ -1, $X_{n2} = $ -1, $X_{n1}X_{n2} = 1$ if in group 4 (black female). When the general form of Eq. (3) is applied, one has:

$$\begin{aligned} \log\left(\frac{P_n}{1-P_n}\right) &= \tau_0 + \tau_1 T_n + \tau_{21} X_{n1} + \tau_{22} X_{n2} + \tau_{23} X_{n1} X_{n2} \\ &\quad + \tau_{31} X_{n1} T_n + \tau_{32} X_{n2} T_n + \tau_{33} X_{n1} X_{n2} T_n, \end{aligned} \tag{9}$$

in which $\boldsymbol{\tau}_2{}' = (\tau_{21},\ \tau_{22},\ \tau_{23})$, $\boldsymbol{\tau}_3{}' = (\tau_{31},\ \tau_{32},\ \tau_{33})$, and $\mathbf{X}_n{}' = (X_{n1},\ X_{n2},\ X_{n1}X_{n2})$. With the design matrix in Eq. (8), τ_{21} depicts the main effect of factor A on uniform DIF, τ_{22} depicts the main effect of factor B on uniform DIF, τ_{23} depicts the interaction effect of factors A and B on uniform DIF, τ_{31} depicts the main effect of factor A on nonuniform DIF, τ_{32} depicts the main effect of factor B on nonuniform DIF, and τ_{33} depicts the interaction effect of factors A and B on nonuniform DIF. When there is no nonuniform DIF, Eq. (9) becomes

$$\text{White Male } (X_1{=}1,\ X_2{=}1,\ X_1X_2{=}1):\ \log\left(\frac{P_n}{1-P_n}\right)=\tau_0+\tau_1T_n+\tau_{21}+\tau_{22}+\tau_{23}, \tag{10}$$

$$\text{White Female } (X_1{=}-1,\ X_2{=}1,\ X_1X_2{=}-1):\log\left(\frac{P_n}{1-P_n}\right)=\tau_0+\tau_1T_n-\tau_{21}+\tau_{22}-\tau_{23}, \tag{11}$$

$$\text{Black Male } (X_1{=}1,X_2{=}-1,X_1X_2{=}-1):\log\left(\frac{P_n}{1-P_n}\right)=\tau_0+\tau_1T_n+\tau_{21}-\tau_{22}-\tau_{23}, \tag{12}$$

$$\text{Black Female } (X_1{=}-1,X_2{=}-1,\ X_1X_2{=}1):\log\left(\frac{P_n}{1-P_n}\right)=\tau_0+\tau_1T_n-\tau_{21}-\tau_{22}+\tau_{23}, \tag{13}$$

If $\boldsymbol{\tau}_2{}' = (\tau_{21},\ \tau_{22},\ \tau_{23}) = (0.4,\ -0.3,\ 0.2)$, then it can be shown that, on average, males have a logit 0.8 higher than that of females; white people have a logit 0.6 lower than that of black people; and white males and black females have a logit 0.4 higher than that of white females and black males. A similar interpretation applies to $\boldsymbol{\tau}_3$.

The use of design matrices like Eq. (8) enables users to decompose uniform DIF and nonuniform DIF into a main effect of factor A, a main effect of factor B, and an interaction effect between factors A and B. Furthermore, Eq. (9) can be easily generalized to cover more than two grouping factors, which can be categorical or continuous, as in factorial ANOVA or ANCOVA (analysis of covariance).

The likelihood ratio test can be adopted to statistically test whether the $\boldsymbol{\tau}_2$ and $\boldsymbol{\tau}_3$ vectors are zero. By comparing the likelihood ratio of Eqs. (14) and (3), one can test whether the studied item has DIF:

$$\log\left(\frac{P_n}{1-P_n}\right) = \tau_0 + \tau_1 T_n, \tag{14}$$

against a chi-square distribution with degrees of freedom of the length of $\boldsymbol{\tau}_2$ and $\boldsymbol{\tau}_3$. Likewise, one can compare the likelihood ratio of Eqs. (15) and (3) to test whether the studied item has nonuniform DIF:

$$\log\left(\frac{P_n}{1-P_n}\right) = \tau_0 + \tau_1 T_n + \boldsymbol{\tau}_2{}'\mathbf{X}_n, \tag{15}$$

against a chi-square distribution with degrees of freedom of the length of $\boldsymbol{\tau}_3$. When $\boldsymbol{\tau}_3$ is a zero vector, it is desirable to test whether this item has uniform DIF, which can be done by comparing the likelihood ratio of Eqs. (14) and (15) against a chi-square distribution with degrees of freedom of the length of $\boldsymbol{\tau}_2$. All these equations and likelihood ratio tests can be easily implemented on commercial programs such as SPSS and SAS, or free software such as R.

In the following simulation studies, we were particularly interested in two questions: (a) Could the FLR method detect uniform DIF effectively under different conditions, as compared to traditional LR methods? and (b) Could the FLR method detect nonuniform DIF effectively under different conditions, as compared to traditional LR methods? Each question was answered by a simulation study. In both simulation studies, there were two grouping variables and each had two levels.

15.2 Simulation Study 1: Uniform DIF

15.2.1 Design

Let the two grouping variables be denoted A and B. Let X_1 be the dummy variable for factor A, X_2 be the dummy variable for factor B, and X_1X_2 be the dummy variable for factors A and B. The design matrix was identical to that in Eq. (5). Each of the four groups of examinees had a sample size of 125, and their ability levels were generated from $N(0, 1)$. There were 21 items in the test, in which items 1–20 were treated as an anchored test to place all the examinees from different groups on the same scale, so that item 21 could be detected for DIF. The item responses followed the Rasch model. There were three independent variables: (a) percentage of DIF items in the anchored test, 0, 10, and 20 % DIF items in the 20-item anchored test; (b) DIF size in the studied item, 0, 0.2, 0.4, and 0.6 logits; and (c) DIF source, consisting of main effect of factor A, main effects of factors A and B, the interaction effect, main effect of factor A and the interaction effect, and main effects of factors A and B and the interaction effect. Let the difficulty parameter be b when an item did not have DIF. It became $b \pm 0.2$, $b \pm 0.4$, and $b \pm 0.6$ for the four groups, according to the design matrix in Eq. (5) when the DIF size was 0.2, 0.4, and 0.6, respectively. Although an anchored test should preferably include exclusively DIF-free items, in reality, DIF items may be included in an anchored test. Inclusion of DIF items often results in poorer DIF assessment (Narayanan and Swaminathan 1996; Rogers and Swaminathan 1993). Scale purification procedures for logistic regression methods have been developed (French and Maller 2007). However, this study did not consider scale purification because its major purpose was to evaluate the FLR method and others, even when the anchored test included DIF items.

A total of 76 conditions were examined with 1,000 replications under each condition. Each simulated dataset was analyzed with the following four methods:

1. The LR-A method in which DIF analysis was conducted to assess DIF of grouping variable A;
2. The LR-B method in which DIF analysis was conducted to assess DIF of grouping variable B;
3. The LR-AB method in which DIF analysis was conducted to assess DIF of grouping variables A and B consecutively; and
4. The proposed FLR method.

Although there were two grouping variables and DIF analysis should be conducted on both variables (meaning that the LR-A and LR-B methods were not applicable in practice), the LR-A and LR-B methods were adopted, by which the LR-AB and FLR methods can be compared. The nominal level of hypothesis testing was set at 0.05. Note that in the LR-AB method there were two hypothesis tests, so the Bonferroni adjustment was applied.

The outcome variables were the Type I error rate and the power rate. The empirical Type I error rate (false positive rate) was computed as how many times in the 1,000 replications a DIF-free studied item (DIF size = 0) was mistakenly declared as having DIF; and the empirical power rate (true positive rate) was computed as how many times in the 1,000 replications a DIF item was correctly detected as having DIF.

It was expected that (a) when the anchored tests did not contain any DIF items, all four methods would yield well-controlled Type I error rates; (b) when the anchored tests contained DIF items, the performance of these four methods would be degraded; (c) the FLR method would have higher power than the other methods when the DIF source contained the interaction of factors A and B; and (d) the larger the DIF size, the higher the power rate.

15.2.2 Results

15.2.2.1 Empirical Type I Error Rates

When the anchored test did not contain any DIF items, the empirical Type I error rates were 0.058, 0.058, 0.053, and 0.047 for the FLR, LR-AB, LR-A, and LR-B methods, respectively. All methods yielded well-controlled Type I error rates, as expected. When the anchored test contained 10 % DIF items, as shown in the upper panel of Table 15.1, the Type I error rates were inflated, especially when the DIF size was large. In addition, it was evident that the LR-AB and FLR methods were more adversely affected than the LR-A and LR-B methods by the inclusion of DIF items in the anchored test. When the anchored test contained 20 % DIF items, as shown in the lower panel of Table 15.1, the inflation in the Type I error rates was even worse than it was in the condition of 10 % DIF items. For example, when the DIF source contained the interaction between factors A and B and the DIF size was

large, the FLR method yielded a Type I error rate of 0.077 when there were 10 % DIF items in the anchored test, and 0.235 when there were 20 % DIF items. Thus, the second expectation was supported, too.

15.2.2.2 Empirical Power Rates

First, consider the case where the anchored test did not contain any DIF items. As shown in the upper panel of Table 15.2, when the DIF source contained exclusively the interaction between factors A and B, only the FLR method yielded high power rates: 0.462, 0.971, and 1.000 when the DIF size was small (0.2 logits), medium (0.4 logits), and large (0.6 logits), respectively, whereas the other three methods yielded power rates between 0.033 and 0.050. A close inspection of the panel revealed that the FLR method substantially outperformed the other three methods as long as the DIF source contained the interaction. When the DIF source contained exclusively the main effect of factor A, the LR-A method had the highest power rates, and the LR-B had the lowest power rates. It was also very clear that the larger the DIF size, the higher the power rate.

Second, consider the case in which the anchored test contained 10 or 20 % (uniform) DIF items, as shown in the middle and lower panels. Take the power rates when the anchored tests did not contain any DIF items as a reference. Across the 15 conditions (5 DIF sources by 3 DIF sizes), the mean power rate was increased by 1, 2, 5, and 2 %, for the FLR, LR-AB, LR-A, and LR-B methods, respectively, when the anchored tests contained 10 % DIF items, and increased by 4, -5, -4, and 2 % for the four methods, respectively, when the anchored tests contained 20 % DIF items. It appears that the inclusion of 10 or 20 % (uniform) DIF items in the anchored test did not substantially affect the power rates of these four methods.

15.3 Simulation Study 2: Nonuniform DIF

15.3.1 Design

This simulation study focused on the assessment of nonuniform DIF. Item responses were simulated according to the three-parameter logistic model. The settings were identical to those in Simulation Study 1, except (a) the discrimination parameters were generated from a log-normal distribution with mean of 0 and variance of 0.1, and the guessing parameters were fixed as 0.2 for all items; (b) the DIF occurred only on the discrimination parameters across different groups of examinees, and the DIF size on a logarithm scale was set at 0, 0.13, 0.26, and 0.39, representing DIF-free, small, medium, and large DIF effects, respectively. Let the discrimination parameter be a when an item did not have DIF. It became

Table 15.1 Type I error rates ($^{0}/_{00}$) of the four methods in uniform DIF

	FLR			LR-AB			LR-A			LR-B		
DIF source	Small	Medium	Large	Small	Medium	Large	Small	Medium	Large	Small	Medium	Large
10 % DIF items												
Interaction	49	81	77	53	60	48	45	57	53	44	53	45
Main effect of A	61	65	88	59	66	94	69	73	103	48	57	60
Main effect of A and interaction	63	80	133	57	73	108	59	77	120	60	45	58
Main effects of A and B	61	67	122	58	74	143	43	66	126	59	65	136
Main effects of A and B and interaction	70	75	155	60	63	133	57	62	114	58	75	121
20 % DIF items												
Interaction	73	115	235	52	43	61	52	58	53	46	40	43
Main effect of A	69	128	155	74	141	180	89	182	216	54	42	54
Main effect of A and interaction	74	132	379	60	90	220	68	111	291	46	54	60
Main effects of A and B	62	192	411	78	219	418	83	173	327	66	168	338
Main effects of A and B and interaction	77	220	616	80	177	404	84	150	321	66	152	347

Note: Small, medium, and large refer to DIF effect size

Table 15.2 Power rates ($^{0}/_{00}$) of the four methods in uniform DIF

	FLR			LR-AB			LR-A			LR-B		
DIF source	Small	Medium	Large	Small	Medium	Large	Small	Medium	Large	Small	Medium	Large
0 % DIF items												
Interaction	462	971	1,000	45	33	49	47	49	50	46	43	44
Main effect of A	559	1,000	999	618	1,000	1,000	707	1,000	1,000	50	44	42
Main effect of A and interaction	929	1,000	1,000	700	998	1,000	782	999	1,000	63	60	122
Main effects of A and B	803	1,000	1,000	789	998	1,000	662	986	1,000	626	982	1,000
Main effects of A and B and interaction	642	1,000	1,000	429	993	1,000	352	904	1,000	319	919	1,000
10 % DIF items												
Interaction	499	948	1,000	43	43	41	42	40	47	45	36	44
Main effect of A	499	959	996	556	977	998	657	985	1,000	31	47	41
Main effect of A and interaction	836	1,000	1,000	593	994	1,000	691	999	1,000	57	84	417
Main effects of A and B	821	1,000	1,000	808	1,000	1,000	653	998	1,000	660	1,000	1,000
Main effects of A and B and interaction	958	1,000	1,000	836	1,000	1,000	724	977	1,000	713	980	1,000
20 % DIF items												
Interaction	364	949	999	41	58	51	42	62	54	47	47	43
Main effect of A	346	976	1,000	385	989	1,000	479	993	1,000	63	42	49
Main effect of A and interaction	766	991	1,000	510	858	994	607	921	999	48	141	547
Main effects of A and B	608	998	1,000	580	999	1,000	464	969	1,000	457	961	1,000
Main effects of A and B and interaction	749	999	1,000	482	969	1,000	392	884	1,000	403	873	1,000

Note: Small, medium, and large refer to DIF effect size

$\log(a) \pm 0.13$, $\log(a) \pm 0.26$, $\log(a) \pm 0.39$, for the last three groups according to the design matrix in Eq. (8) when the DIF size was 0.13, 0.26, and 0.39, respectively. Note that the difficulty parameter did not exhibit DIF.

15.3.2 Results

15.3.2.1 Empirical Type I Error Rates

The Type I error rates were 0.054, 0.048, 0.052, and 0.044 for the FLR, LR-AB, LR-A, and LR-B methods, respectively, suggesting a very good control. As shown in Table 15.3, when the anchored test contained 10 or 20 % DIF items, the Type I error rates for the four methods were still very close to their expected value of 0.05. A comparison of the Type I error rates in Tables 15.1 (uniform DIF) and 15.3 (nonuniform DIF) reveals that the inclusion of uniform DIF items (with difference in the difficulty parameters across groups) in the anchored test had a more adverse effect on the DIF assessment than the inclusion of nonuniform DIF items (with difference in the discrimination parameters across groups). This was mainly because the inclusion of uniform DIF items in the anchored test would deteriorate the correspondence between the raw test score used in the LR methods and the ability level simulated from IRT models, whereas the correspondence was not substantially affected by the inclusion of nonuniform DIF items. Note that including DIF items with difference in both the difficulty and discrimination parameters across groups (referred to as nonuniform DIF items in the literature) would also exhibit an adverse effect.

15.3.2.2 Empirical Power Rates

The upper panel of Table 15.4 shows the power rates of the four methods when the anchored test did not contain any DIF items. When the DIF source contained exclusively the interaction between factors A and B, only the FLR method yielded high power rates: 0.084, 0.186, and 0.538 when the DIF size on the discrimination parameter was small (0.13), medium (0.26), and large (0.39), respectively; whereas the other three methods yielded power rates between 0.036 and 0.055. The panel also shows that the FLR method substantially outperformed the other three methods as long as the DIF source contained the interaction. When the main effect of factor was the only DIF source, the LR-A method had the highest power rates, and the LR-B had the lowest power rates. Furthermore, the larger the DIF size, the higher the power rate.

The middle and lower panels of Table 15.4 show the power rates of the four methods where the anchored test contained 10 or 20 % (nonuniform) DIF items, respectively. Take the power rates when the anchored tests did not contain any DIF items as a reference. Across the 15 conditions (5 DIF sources by 3 DIF sizes), the

Table 15.3 Type I error rates ($^{0}/_{00}$) of the four methods in nonuniform DIF

	FLR			LR-AB			LR-A			LR-B		
DIF source	Small	Medium	Large	Small	Medium	Large	Small	Medium	Large	Small	Medium	Large
10 % DIF items												
Interaction	45	58	54	37	58	47	44	60	52	35	45	42
Main effect of A	44	54	48	48	46	39	45	51	46	47	49	40
Main effect of A and interaction	54	47	51	61	49	48	61	56	44	43	43	47
Main effects of A and B	64	40	46	65	42	49	61	44	45	58	41	44
Main effects of A and B and interaction	47	49	61	50	46	58	55	48	51	43	50	51
20 % DIF items												
Interaction	53	41	58	40	50	57	41	56	50	33	40	56
Main effect of A	54	45	56	52	49	53	60	37	72	54	55	44
Main effect of A and interaction	50	53	46	44	48	59	50	46	51	48	48	59
Main effects of A and B	50	53	55	39	60	56	53	59	45	46	52	51
Main effects of A and B and interaction	53	41	58	40	50	57	41	56	50	33	40	56

Note: Small, medium, and large refer to DIF effect size

Table 15.4 Power rates ($^{0}/_{00}$) of the four methods in nonuniform DIF

	FLR			LR-AB			LR-A			LR-B		
DIF source	Small	Medium	Large	Small	Medium	Large	Small	Medium	Large	Small	Medium	Large
0 % DIF items												
Interaction	84	186	538	44	43	54	47	43	55	43	36	49
Main effect of A	117	573	505	128	665	596	165	751	677	45	49	49
Main effect of A and interaction	200	244	859	138	148	621	194	192	709	46	65	46
Main effects of A and B	201	379	794	202	389	747	154	300	616	170	310	629
Main effects of A and B and interaction	183	699	1,000	125	443	901	110	351	719	112	321	709
10 % DIF items												
Interaction	87	319	465	45	46	49	60	36	45	46	51	50
Main effect of A	81	225	943	85	276	969	104	348	985	59	50	42
Main effect of A and interaction	132	473	620	97	274	274	127	345	347	49	55	40
Main effects of A and B	97	353	763	116	356	749	95	283	613	103	283	576
Main effects of A and B and interaction	151	750	853	115	442	592	104	336	456	102	334	484
20 % DIF items												
Interaction	287	688	768	54	66	52	50	64	53	50	50	52
Main effect of A	52	168	464	53	198	548	59	276	646	49	43	58
Main effect of A and interaction	114	312	524	69	180	325	103	247	397	38	43	67
Main effects of A and B	205	987	809	198	980	785	168	917	630	173	912	638
Main effects of A and B and interaction	145	514	749	101	354	523	82	267	406	100	282	404

Note: Small, medium, and large refer to DIF effect size

mean power rate was increased by -2, -5, -5, and -2 % for the FLR, LR-AB, LR-A, and LR-B methods, respectively, when the anchored tests contained 10 % DIF items, and increased by 1, -5, 2, and -5 % for the four methods, respectively, when the anchored tests contained 20 % DIF items. Thus it can be concluded that the inclusion of 10 or 20 % nonuniform DIF items in the anchored test did not substantially affect the Type I error rates or power rates of these four methods.

Conclusion and Discussion

DIF assessment may be conducted across several grouping factors. In addition to detecting whether an item has DIF, it is also informative to account for DIF source: whether the DIF came from a specific grouping factor or from their interactions. In this study, we incorporated a factorial procedure on the commonly used logistic regression method. The use of design matrices, like those commonly used in factorial ANOVA, enables the decomposition of DIF source into main effects of individual grouping factors and their interaction effects. The parameters in the FLR methods can be interpreted as they are in standard logistic regression. Furthermore, being a nonparametric method, the FLR method is simple to implement and fast to converge, and does not require specification of an item response model or a large sample.

Two simulation studies were conducted to evaluate the performance of the FLR in the detection of uniform and nonuniform DIF, as compared to three other LR methods. The simulation results demonstrate the superiority of the FLR method over the LR-A, LR-B, and LR-AB methods when there was an interaction effect between grouping factors. In reality, interactions among grouping factors can occur and their magnitude may be too large to neglect. In such cases, among the four methods investigated in this study, only the FLR method can yield a higher power of detection. We also investigated whether the FLR method would be adversely affected by including 10 or 20 % DIF items in the anchored test. The results showed a small deflation in the mean power rates, but a substantial inflation in Type I error rates when the anchored test had uniform DIF items with large DIF sizes. The adverse effect was less obvious when the DIF items in the anchored test had different discrimination parameters but the same difficulty parameters across groups.

In this study, all groups were simulated to have an equal mean ability (i.e., no impact). In reality, different groups may have different means (i.e., with impact). It has been shown that the LR method yields inflated Type I error rates and deflated power rates when there is a large impact (Bolt and Gierl 2006; Güler and Penfield 2009). The test raw scores do not match ability levels and thus, the approach fails to place different groups on the same scale for DIF assessment, when groups have very different means. Roussos and Stout (1996) suggest a longer anchored test for large impacts. Even so, the

(continued)

advantages of the FLR method over the LR method would remain unchanged even with large impacts.

This study has implications for DIF research methodology and enables practitioners to assess DIF sources for future item revision. The FLR method can be generalized to assess DIF in polytomous items. Future studies can evaluate the FLR method under different conditions of test lengths, sample sizes, and combinations of uniform and nonuniform DIF items. It is also important to evaluate the FLR method when there is an impact, or when tests consist of both dichotomous and polytomous items.

Acknowledgment The research was supported by the General Research Fund, Hong Kong Research Grants Council (No. 844110).

References

Bolt D, Gierl MJ (2006) Testing features of graphical DIF: application of a regression correction to three nonparametric statistical tests. J Educ Meas 43:313–333. doi:10.1111/j.1756-3984.2006.00019.x

Candell GL, Hulin CL (1986) Cross-language and cross-cultural comparisons in scale translations: independent sources of information about item nonequivalence. J Cross Cult Psychol 17:417–440. doi:10.1177/0022002186017004003

Chen H-F, Jin K-Y, Wang W-C (2012) Assessing differential item functioning when interactions among subgroups exist. Paper presented at the Taiwan education research association international conference on education, Kaohsiung, Taiwan

DeMars CE (2010) Type I error inflation for detecting DIF in the presence of impact. Educ Psychol Meas 70:961–972. doi:10.1177/0013164410366691

Drasgow F (1987) Study of the measurement bias of two standardized psychological tests. J Appl Psychol 72:19–29. doi:10.1037/0021-9010.72.1.19

French BF, Maller SJ (2007) Iterative purification and effect size use with logistic regression for differential item functioning detection. Educ Psychol Meas 67:373–393

Güler N, Penfield RD (2009) A comparison of the logistic regression and contingency table methods for simultaneous detection of uniform and nonuniform DIF. J Educ Meas 46:314–329. doi:10.1111/j.1745-3984.2009.00083.x

Iwata N, Turner RJ, Lloyd DA (2002) Race/ethnicity and depressive symptoms in community-dwelling young adults: a differential item functioning analysis. Psychiatry Res 110:281–289. doi:10.1016/S0165-1781(02)00102-6

Kim J, Oshima TC (2013) Effect of multiple testing adjustment in differential item functioning detection. Educ Psychol Meas 73:458–470. doi:10.1177/0013164412467033

Kim SH, Cohen AS, Park TH (1995) Detection of differential item functioning in multiple groups. J Educ Meas 32:261–276. doi:10.1111/j.1745-3984.1995.tb00466.x

Li YJ, Brooks GP, Johanson GA (2012) Item discrimination and Type I error in the detection of differential item functioning. Educ Psychol Meas 72:847–861. doi:10.1177/0013164411432333

Narayanan P, Swaminathan H (1994) Performance of the Mantel–Haenszel and simultaneous item bias procedures for detecting differential item functioning. Appl Psychol Meas 18:315–328. doi:10.1177/014662169401800403

Narayanan P, Swaminathan H (1996) Identification of items that show nonuniform DIF. Appl Psychol Meas 20:257–274. doi:10.1177/014662169602000306

Penfield RD (2001) Assessing differential item functioning among multiple groups: a comparison of three Mantel–Haenszel procedures. Appl Meas Educ 14:235–259. doi:10.1207/S15324818AME1403_3

Rogers HJ, Swaminathan H (1993) A comparison of logistic regression and Mantel–Haenszel procedures for detecting differential item functioning. Appl Psychol Meas 17:105–116. doi:10.1177/014662169301700201

Roussos L, Stout W (1996) A multidimensionality-based DIF analysis paradigm. Appl Psychol Meas 20:355–371

Somes GW (1986) The generalized Mantel–Haenszel statistics. Am Stat 40:106–108. doi:10.1080/00031305.1986.10475369

Swaminathan H, Rogers HJ (1990) Detecting differential item functioning using logistic regression procedures. J Educ Meas 27:361–370. doi:10.1111/j.1745-3984.1990.tb00754.x

Wang W-C (2000a) Modeling effects of differential item functioning in polytomous items. J Appl Meas 1:63–82

Wang W-C (2000b) The simultaneous factorial analysis of differential item functioning. Methods Psychol Res 5:56–76

Zwick R, Donoghue JR, Grima A (1993) Assessment of differential item functioning for performance tasks. J Educ Stat 15:185–187. doi:10.1111/j.1745-3984.1993.tb00425.x