Introduction

Differential item functioning (DIF) is said to be present when examinees from two groups which have the same amount of the ability measured by the test perform differently on a test item. DIF analysis of items in the dichotomous and polytomous cases have both been studied. Extensive reviews of these DIF methodologies can be seen in Millsap and Everson (1993), Camilli and Shepard (1994), Potenza and Dorans (1995), and Penfield and Lam (2000).

In general, parametric DIF procedures could be classified into three classes of approaches. The first class is to compare item parameters estimated from the two groups of examinees, such as Lord's χ^2 (1980). The second one is to compare areas between the item characteristic curves or the expected item score functions estimated from the two groups (e.g., Kim & Cohen, 1991; Cohen et al., 1993; Oshima et al., 1994; Flowers et al., 1999). The last class is to compare the likelihood functions, as in the Likelihood Ratio test (LR; Thissen, Steinberg, & Gerard, 1986) and the logistic regression procedure (Logi; Swaminathan & Rogers, 1990; French & Miller, 1996). In addition, there are also nonparametric DIF procedures which do not assume a specific statistical model, such as the Mantel-Haenszel procedure (MH; Holland & Thayer,1988; Zwick et al., 1993), the simultaneous item bias procedure (SIBTEST; Shealy & Stout,1993) and its polytomous extension (Poly-SIBTEST; Chang, Mazzeo, & Roussos,1996).

Logi was proposed by Swaminathan and Rogers (1990) in the first place for dichotomously scored items, and was later extended to polytomous items by recoding data into multiple

dichotomies (Miller & Spray, 1993; French & Miller, 1996). French and Miller (1996) compared Logi in three different coding schemes – continuation ratio logits, cumulative logits, and adjacent categories logits model and found the continuation and cumulative logits to be powerful in detecting most forms of DIF in large samples.

LR was first proposed for DIF analysis by Thissen, Steinberg, and Gerard (1986). Kim and Cohen (1998) investigated its performance under different sample sizes and ability distributions using the graded response model (GRM; Samejima, 1969). It was shown that the Type I error rates of LR under all conditions were very close to those expected at most of the α levels under consideration.

Raju et al. (1995) presented DFIT and offered an empirical demonstration using dichotomous data. It was extended to the dichotomous multidimensional case by Oshima et al. (1997), and to the polytomous unidimensional case by Flowers et al. (1999). Its effectiveness in identifying DIF items have been shown in the above studies.

In a comparison to MH (Holland & Thayer,1988) and SIBTEST (Shealy & Stout, 1993) for dichotomous items, Logi was found as effective as MH and SIBTEST in detecting uniform DIF (Narayanan & Swaminathan, 1994; Rogers & Swaminathan, 1993). In the detection of nonuniform DIF, Logi was shown more powerful than MH and as powerful as SIBTEST (Narayanan & Swaminathan, 1996). The notions of uniform and nonuniform DIF would be reviewed later in the section of DIF indices.

Bolt (2002) compared the performance of LR, DFIT and Poly-SIBTEST procedures under GRM (Samejima, 1969) in terms of their Type I error and Power. When data were

generated from GRM, LR was superior to DFIT in its Type I error and Power. LR and DFIT were both better than Poly-SIBTEST with small sample sizes.

As far as we are aware, Logi has not yet been compared to either LR or DFIT in the literature. Logi has three major advantages. First, it could treat the covariate variable as continuous without losing power for stratification. Secondly, it can be implemented easily in practice. Thirdly, it could tell the kind of DIF, i.e., uniform or nonuniform, an item exhibits. To evaluate Logi in comparison to LR and DFIT, we investigate the performance of the Logi, LR, and DFIT procedures through a simulation study using GRM.

Lastly, in order to see to what extent the results of the DIF analyses of real data might differ using different indices, we analyze test items in the Chinese Version 2.0 of the Cognitive Ability Screening Instrument (CASI C-2.0) with the three procedures and compare the results with those obtained in a previous study by Crane et al. (2004).

DIF Indices

In most cases, there are two groups under investigation in a DIF analysis. The two groups are called the reference group and the focal group. Conventionally, the reference group is the group taken as a standard for the group of interest, i.e., the focal group, to be compared to. Generally speaking, DIF detection mainly aims to test the null hypothesis of

H_0: the item functions equally for the reference and focal groups (No DIF),

against the alternative hypothesis of

H_A: the item functions unequally for the reference and focal groups (DIF).

As mentioned above, Logi has been said to be superior to MH procedure in detecting DIF of both the dichotomous and polytomous items (Swaminathan & Rogers, 1990; French & Miller, 1996). LR and DFIT have also been shown to perform better than Poly-SIBTEST (Bolt, 2002). Therefore, here we mainly compare the performance of the three DIF indices, Logi, LR, and DFIT, under the GRM condition. We now describe first the graded response model (GRM).

Graded Response Model (GRM)

Item response models are mathematical functions which relate the category response probability to both the examinees' abilities and the parameters that characterize the item. There are many IRT models for polytomous items which are also commonly used especially in performance assessment. Just to name a few, there are Samejima's (1969) graded response model (GRM), Andersen's (1977) rating scale model, Master's (1982) partial credit model (PCM), and the generalized partial credit model (Muraki, 1992). Among them, the majority of DIF studies in the literature have been done using GRM, partly because the softwares are easily available for its estimation and testing.

GRM deals with ordinally scored polytomous items. The ordered categories of the items include rating such as disagree, no comment, and agree, used in attitude surveys, or partial credit given according to the degree of attainment in solving a problem. Under the GRM, the probability of examinee l responding above category k to item i, with possible scoring

$k = 1, 2, ..., m$, is:

$$P_{ik}^{*}(\theta_l) = \frac{\exp[a_i(\theta_l - b_{ik})]}{1 + \exp[a_i(\theta_l - b_{ik})]}, \tag{1}$$

where a_i is the discrimination parameter for item i, b_{ik} is the category threshold parameter associated with item i, and $b_{il} > b_{ik}$ whenever $l > k$, and θ_l is the latent trait level of examinee l. Equation (1) is referred to as the boundary response function (BRF) and the number of the BRFs within each item equals to the number of category threshold parameters.

The probability of scoring in each category is computed from adjacent boundary probabilities. That is,

$$P_{ik}(\theta_l) = P_{i,k-1}^{*}(\theta_l) - P_{ik}^{*}(\theta_l). \tag{2}$$

Equation (2) is referred to as the item category response function (ICRF). For category 1 and m lacking an adjacent category, Samejima (1969) defined $P_{i0}^{*}(\theta_l) = 1$ and $P_{im}^{*}(\theta_l) = 0$. Thus $P_{im}(\theta_l) = P_{i,m-1}^{*}(\theta_l)$ and $P_{i1}(\theta_l) = 1 - P_{i1}^{*}(\theta_l)$.

For each item, the number of category threshold parameters is one less than the number of categories. For example, a four-category item requires four item parameters – three category threshold parameters ($b's$) and one discrimination parameter (a). In addition, there would be three BRFs for this item.

Method 1: Logistic Regression Procedure

Logistic regression procedures have been shown to be useful in detecting uniform and nonuniform DIF of polytomous items (French & Miller, 1996; Zumbo, 1999). Uniform DIF occurs

when the responses of the examinees from the two groups differ uniformly at all trait levels. Conversely, nonuniform DIF exists when there is an interaction between trait level and group assignment. Here we used the logistic regression model in continuation ratio logits, and adopted the same strategy for determining uniform and nonuniform DIF as in Crane et al. (2004).

The logistic regression model used was:

$$f(y_i|\theta_l, g) = \beta_{i0} + \beta_{i1} * \theta_l + \beta_{i2} * g + \beta_{i3}(g * \theta_l), \tag{3}$$

where the left side of the function is the continuation ratio logit, y_i is the examinee's response on item i, θ_l is the ability of examinee l, and g represents his or her group assignment ($g = R$ or F). First of all, we examine the presence of nonuniform DIF by looking for the statistically significant interaction term. That is, if the Bonferroni-adjusted p value associated with the β_{i3} term was less than α, nonuniform DIF was detected for item i. If not, the interaction term would be eliminated for the following analysis.

For items free of nonuniform DIF, we then compare the β_1 terms of the models with and without group assignment. If a 10% difference between the β_1 coefficients with and without group assignment was found in the model, uniform DIF was said to be present (Maldonado & Greenland, 1993). That is, if the ratio of β_1' and β_1 in the following two models:

$$f(y|\theta, g) = \beta_0' + \beta_1' * \theta + \beta_2' * g$$

and

$$f(y|\theta, g) = \beta_0 + \beta_1 * \theta$$

is not within the range from 0.9 to 1.1, the item is said to contain uniform DIF.

Method 2: Likelihood Ratio Test

Suppose that the null hypothesis is $H_0 : \gamma = Set_C$ (where Set_C contains N parameters) and the alternative hypothesis is $H_A : \gamma = Set_A$ (where Set_A contains N+M parameters), the likelihood ratio (LR) of interest for the two models is

$$\text{LR} = \frac{L^*(\text{Model C})}{L^*(\text{Model A})}, \tag{4}$$

where L^*(Model) refers to the maximum value of the likelihood function L(Model). Define twice the negative natural log transformation of LR as

$$G^2 = (-2)\ln(\text{LR}) = [(-2)\ln L^*(\text{Model C})] - [(-2)\ln L^*(\text{Model A})]. \tag{5}$$

In large samples, G^2 is approximately distributed as a χ^2 distribution with degrees of freedom M.

In evaluating the differences between item responses from the two groups in a DIF analysis under GRM, the compact model of the H_0 hypothesis is defined by constraining all item parameters to be equal for the reference and the focal groups, whereas in the augmented model of the H_A hypothesis, only the parameters of the studied item are allowed to differ across groups. If the statistic G^2 exceeds a critical χ^2 value at a given level of α, the null hypothesis of no DIF is rejected.

Method 3: Differential Functioning of Items and Tests Procedure

If two items have the same expected scores (ES) or item response functions (IRFs) in the GRM, they would have the same number of scoring categories and the same item category response functions (ICRFs) (Chang & Mazzeo, 1994). Conversely, if two items have different ICRFs, they would result in different expected scores. Therefore, item i is theoretically considered to have DIF if, for any trait level θ_l,

$$\mathrm{ES}_{iR}(\theta_l) \neq \mathrm{ES}_{iF}(\theta_l),$$

where $\mathrm{ES}_{iR}(\theta_l)$ and $\mathrm{ES}_{iF}(\theta_l)$ are respectively the expected scores of item i for an examinee with ability θ_l in the reference and the focal groups. That is,

$$\mathrm{ES}_{ig}(\theta_l) = \sum_{k=1}^{m} kP_{ik,g}(\theta_l), \quad g = R \text{ or } F.$$

The difference between the two expected scores, defined as $d_i(\theta_l) = \mathrm{ES}_{iF}(\theta_l) - \mathrm{ES}_{iR}(\theta_l)$, could therefore be used as a measure of DIF.

Moreover, Raju et al (1995) considered a measure for detecting the bias at the test level, the differential test functioning (DTF), with respect to examinees with ability θ_l as

$$D^2_{\theta_l} = [T_F(\theta_l) - T_R(\theta_l)]^2,$$

where $T_F(\theta_l)$ is the expected test scores for the focal group examinees with ability θ_l and $T_F(\theta_l) = \sum_{i=1}^{n} \mathrm{ES}_{iF}(\theta_l)$ where n is the number of items in the test. Similarly, $T_R(\theta_l)$ is the expected scores for the reference group examinees with the same θ_l and $T_R(\theta_l) = \sum_{i=1}^{n} \mathrm{ES}_{iR}(\theta_l)$. Thus, the greater the difference is between the two expected test scores, the greater the DTF

is. Moreover, DTF across the focal group examinees is defined as:

$$\text{DTF} = \text{E}_F(D_\theta^2) = \int_\theta D_\theta^2 f_F(\theta) d\theta = \sigma_D^2 + \mu_D^2,$$

where $f_F(\theta)$ is the density function of θ for the focal group examinees, and μ_D^2, σ_D^2 are the mean and variance of $D(= \sum_{i=1}^n d_i)$. In addition, a decomposition of DTF into terms of each item as

$$\text{DTF} = \text{E}_F(D_\theta^2) = \text{E}_F[(\sum_{i=1}^n d_i(\theta))^2] = \sum_{i=1}^n [\text{Cov}(d_i, D) + \mu_{d_i}\mu_D]$$

facilitates the definition of CDIF for each item as

$$\text{CDIF}_i = \text{Cov}(d_i, D) + \mu_{d_i}\mu_D. \tag{6}$$

Under the assumption that all but the studied item are free from DIF, $d_j = 0$ for all $j \neq i$ and $D = d_i$, where i is the item studied, another index, noncompensatory DIF (NCDIF), is defined as

$$\text{NCDIF}_i = \sigma_{d_i}^2 + \mu_{d_i}^2.$$

Or equivalently, for each item i,

$$\text{NCDIF}_i = \text{E}_F[d_i(\theta)^2] = \int_\theta d_i(\theta)^2 f_F(\theta) d\theta, \tag{7}$$

where $f_F(\theta)$ is again the density function of θ for examinees in the focal group.

For CDIF is not as stable an index as NCDIF (Flowers et al., 1999), here we adopted NCDIF as the index for our DIF study. NCDIF is distributed as a χ^2 under a null hypothesis of no DIF. However, the χ^2 distribution for NCDIF was overly sensitive for large sample, hence an empirical critical value has been suggested instead (Flowers et al., 1999). The

empirical critical value is determined from the distribution of NCDIFs generated under no DIF condition by the percentile corresponding to the desired α level.

Simulation Study

The graded response model with 4 response categories was used to generate data in this study. Table 1 displays the item parameters used for the reference group. Items exhibiting DIF were modified by changing the a and/or b parameters of the focal groups, as can be seen in Table 2.

To generate the data, examinee l's ability θ was first randomly drawn from a normal distribution. Based on the simulated θ and the item parameters, we compute the boundary response probabilities $P^*_{i,k}$ for each item i as defined in Equation (1). The examinee's responses on item i is then determined by comparing a uniform random number Y_{li} over [0,1] to those boundary response probabilities such that if

$$P^*_{l,i,k} < Y_{l,i} < P^*_{l,i,k-1}, \tag{8}$$

a category k is assigned as the examinee l's response on item i.

Factors Manipulated

Examinees' responses were simulated under a variety of conditions which may affect the performance of Type I error and Power rates of the DIF procedures. In this study three factors were manipulated with the test length fixed at thirty items.

Table 1: Reference Group Item Parameters

Item	a	b_1	b_2	b_3	Item	a	b_1	b_2	b_3
1	2.06	-0.78	0.04	1.01	16	1.78	-1.68	-0.36	1.20
2	1.27	-1.80	-0.22	1.63	17	2.11	-1.96	0.08	1.12
3	1.82	-1.32	0.29	1.10	18	1.45	-2.18	-0.37	0.75
4	1.66	-1.12	0.10	1.54	19	1.78	-1.68	0.49	1.53
5	1.73	-2.46	0.13	1.94	20	1.74	-0.40	0.01	1.57
6	1.78	-1.30	0.09	1.46	21	1.54	-1.97	0.11	0.68
7	1.67	-0.57	1.46	1.69	22	1.83	-0.60	1.01	2.34
8	1.62	-1.77	0.43	1.30	23	2.10	-0.94	0.72	1.49
9	2.09	-1.83	0.53	1.22	24	2.09	-1.51	0.75	1.96
10	1.31	-0.45	0.53	1.60	25	1.91	-0.14	0.91	1.88
11	1.56	-1.85	-0.49	0.50	26	1.44	-1.86	0.76	1.40
12	1.23	-1.45	0.44	2.19	27	1.88	-0.59	0.00	1.84
13	1.91	-1.51	0.35	0.71	28	1.94	-1.15	0.65	2.72
14	1.55	-1.25	-0.44	1.28	29	1.81	-2.42	0.86	1.92
15	1.47	-0.95	0.74	1.99	30	1.29	-1.00	0.46	2.09

Table 2: Focal Group Item Parameters (Items Not Listed Used the Same Item Parameters as the Reference Group)

Item	a	b_1	b_2	b_3
Condition	1	(7%)		
4	1.67	-0.64	0.20	1.57
6	1.78	-0.30	1.09	2.46
Condition	2	(10%)		
4	1.67	-0.64	0.20	1.57
6	1.78	-0.30	1.09	2.46
9	1.59	-1.83	0.53	1.22
Condition	3	(20%)		
4	1.67	-0.64	0.20	1.57
6	1.78	-0.30	1.09	2.46
9	1.59	-1.83	0.53	1.22
11	1.06	-1.85	0.00	0.50
22	1.82	0.36	1.20	2.35
30	1.29	-1.00	0.76	2.19

First, two sample sizes were considered. 300 examinees were generated for each group as the small size for it is sufficient for stable IRT item parameter estimates. The other sample size of 1000 for each group is used.

Secondly, two different ability distributions were assumed for the focal group. In the "Equivalent" condition (i.e. $d_\theta = 0$), both groups have equal ability distributions of N(0,1). In the "Nonequivalent" condition, the focal group was sampled from an N(-1,1) distribution whereas the reference group was still sampled from the N(0,1) distribution. It resulted in a difference in the mean ability between groups, i.e., $d_\theta = 1$.

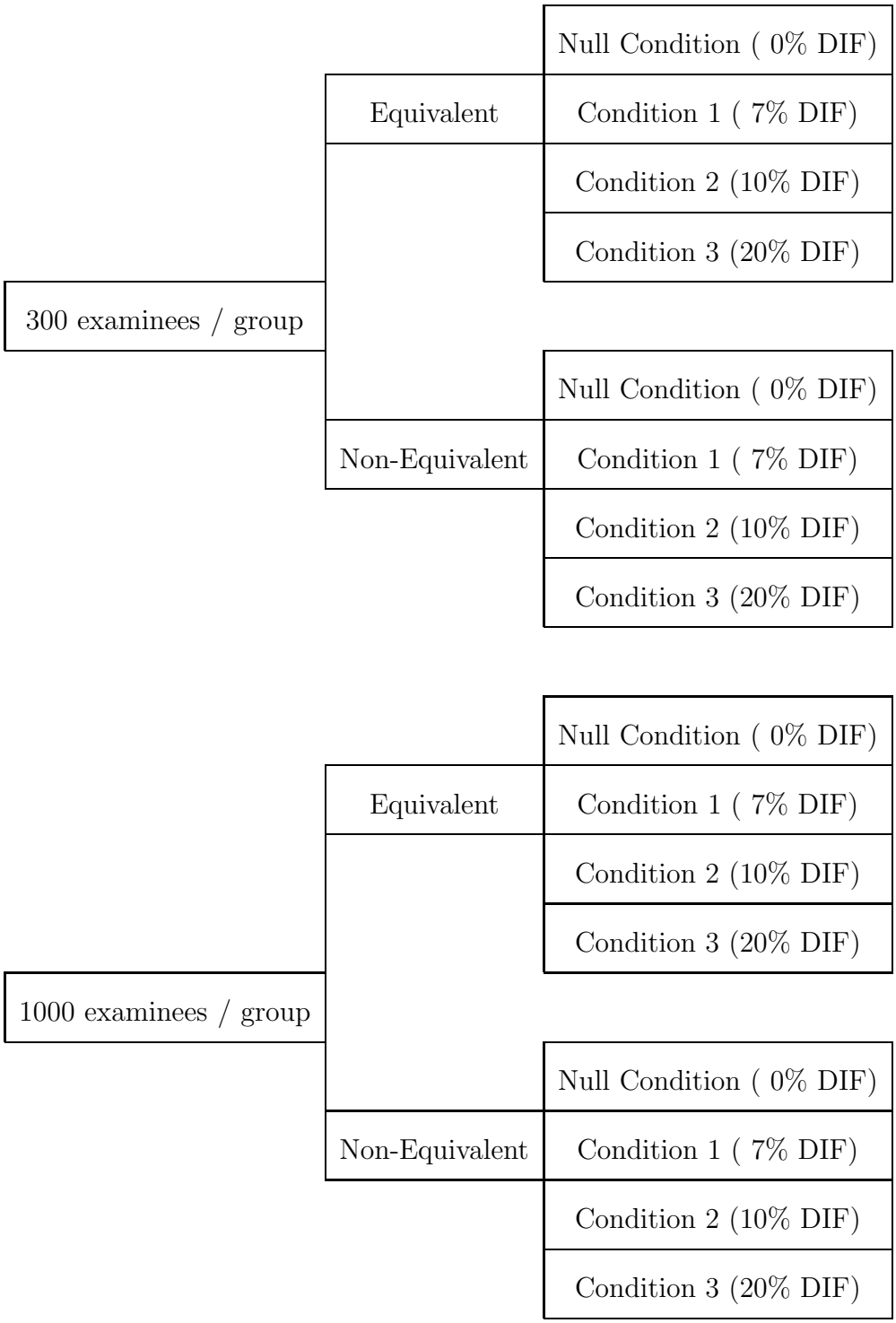
Thirdly, the percentages of items containing DIF were manipulated. Because the percentage of DIF items may contaminate the parameter estimates and affect the Type I error and Power of the DIF procedures. Four levels of percentages (0%,7%,10%,20%) were considered.

Two sample sizes, two ability distributions, and four percentages of DIF item produce the sixteen conditions of data sets. The simulation design is displayed in Table 3. Each condition is replicated 100 times to facilitate Type I error and Power calculations of the three DIF detection procedures.

Logistic Regression Procedure

In the logistic regression procedure, the estimates of examinees' ability $\theta's$ were obtained under GRM with MULTILOG (Thissen,2003). If the interaction term 'β_3' of an item was statistically significant, the item is said to contain nonuniform DIF. In this process, the Bonferroni-adjusted p-value was used. That is, if the p-value for testing β_3 multiplied by 30

Table 3: Simulation Design



was less than $\alpha = 0.05$, the item is said to contain nonuniform DIF.

If the item was lack of nonuniform DIF, we then examine uniform DIF by comparing the β_1 coefficients with and without the group factor in the models. An 10% difference in the coefficients indicate the existence of uniform DIF (Maldonado & Greenland,1993). The analysis of the logistic regression model was done by S-PLUS.

Likelihood Ratio Test

In the LR test, DIF of each item is studied by comparing "NEGATIVE TWICE THE LOG-LIKELIHOOD" of the compact model to that of the augmented model. In the augmented model, only the parameters of the studied item are allowed to differ across groups. A four-category item has four item parameters and consequently the difference of item parameter numbers between the augmented model and the compact model equals to four. Hence, the statistic G^2 is approximately distributed as a χ^2 distribution with four degrees of freedom under the hypothesis of DIF free. So if χ^2 exceeds 9.488, the critical value for $\chi^2(4)$ at $\alpha = 0.05$, the item is judged as exhibiting DIF.

Differential Functioning of Items and Tests Procedure

We first estimated item parameters of each group separately by MULTILOG (Thissen,2003). EQUATE (Baker,1993) was then used to put the two sets of item parameters on the same scale. We computed the NCDIF value of each item using the program written in GAUSS. MULTILOG uses the marginal maximum likelihood estimation, where the ability distribution is assumed to follow $N(0, 1)$. Therefore, in calculating the NCDIF value for each item, we

have for Equation (7),

$$\text{NCDIF}_i = \int_\theta d_i(\theta)^2 f(\theta) d\theta = \int_\theta \frac{1}{\sqrt{2\pi}} d_i(\theta)^2 e^{-\frac{\theta^2}{2}} d\theta,$$

where Gauss-Hermite quadrature method (Stroud & Sechrest, 1966; Bock & Aitkin, 1981) was used to approximate the above integration in our study.

The empirical critical values are determined from the NCDIF distributions of data containing no DIF items. In this study, the chosen critical values correspond to an α of 0.05 in the distribution of the $30 * 100 = 3000$ NCDIF values obtained from the four null conditions. For 300 per group, the critical values for the equivalent and non-equivalent conditions were respectively 0.13429 and 0.088 whereas for 1000 per group, their critical values were 0.039075 and 0.026095.

Results

Type I error study

Table 4 to 6 show the number of times in the 100 replications each item being detected as DIF at $\alpha = 0.05$. The columns are separated according to the factors manipulated in the data: (a) difference of the mean ability distribution across reference and focal groups ($d_\theta = 0$, $d_\theta = 1$), and (b) sample size (300 and 1000 in each group). In the last row we present the mean number of times of detection of all items under each of the four conditions. For example, (Table 4) in Logi with condition of size 1000 and $d_\theta = 1$, item 1 was identified to show DIF twice out of the 100 replications. And for this same condition, each item was in

average indicated as DIF for 2.3 times out of the 100 replications. Because 100 replications were made, the expected number of rejections corresponds to the α of 0.05 was 5 for each item in each condition. Table 7 displayed the summary of mean Type I error rates of the three procedures under the four conditions.

From Table 4 of Logi results, for both sample sizes, the number of rejections was higher in the $d_\theta = 1$ condition than in the $d_\theta = 0$ condition for nearly every item. As a result, we found in the sample of size 300, the Type I error rate in the $d_\theta = 0$ condition (0.13%) was in average lower than that of condition $d_\theta = 1$ (0.76%). In addition, the Type I error rate also increased from 0.20% to 2.30% as d_θ increased to 1000 per group. For each of the d_θ level, the Type I error rates in sample size 1000 level was higher than in 300 level. For example, in the $d_\theta = 0$ condition, the Type I error rate increased from 0.13% to 0.20% as sample size increased from 300 to 1000. To summarize, when the sample size or the mean difference of ability distributions increases, the Type I error rates of Logi would increase. However, with respect to the α of 5%, the Type I error of Logi was relatively low with the maximum of 2.3%.

From Table 5 of LR results, we found that in size of 300, the Type I error rates for $d_\theta = 0$ and $d_\theta = 0$ were both 5.40%. But in size of 1000, the Type I error rate for $d_\theta = 0$ condition (5.83%) was slightly higher than that for $d_\theta = 1$ condition (5.43%). In contrast to Logi, only half of the items satisfied having greater number of rejections in the $d_\theta = 1$ condition than in the $d_\theta = 0$ condition. That is, such a property no longer holds for LR. Moreover, with the same d_θ value, the Type I error rates increased a little when the sample size got larger.

For instance, in the $d_\theta = 0$ condition, the rates increased from 5.40% to 5.83% as sample size raised from 300 to 1000. In total, the Type I error rates of LR, ranging from 5.40% to 5.83%, were consistent with the nominal Type I error rates, as found in Kim and Cohen (1998).

Results for DFIT are displayed in Table 6. Since the critical values were determined from the empirical data, we can see that the mean Type I error rates was automatically constrained at 5% which in turn makes the comparison of DFIT to other methods with respect to their Type I errors somewhat ambiguous.

In short, among the Type I error performance of the three procedures, Logi has the lowest rates. The rates of LR are a little bit higher than those of DFIT. However, their Type I error rates at the test level are all near or below the nominal Type I error rate 5%, for the maximums being 2.30% of Logi, 5.83% of LR, and 5.00% of DFIT.

Table 4: Results of Logi (Number of Rejections Out of 100 Trials) for Sample Sizes 300 and 1000 under null conditions

	$d_\theta = 0$		$d_\theta = 1$			$d_\theta = 0$		$d_\theta = 1$	
Item	300	1000	300	1000	Item	300	1000	300	1000
1	0	0	0	2	16	0	1	2	1
2	0	0	0	2	17	0	0	0	1
3	0	0	0	1	18	0	0	0	0
4	0	0	1	1	19	0	0	0	4
5	1	0	1	1	20	1	0	1	1
6	0	1	0	3	21	0	1	1	2
7	0	1	2	3	22	0	0	2	5
8	1	0	1	3	23	0	0	0	6
9	1	0	0	4	24	0	0	1	1
10	0	0	1	1	25	0	0	2	2
11	0	0	0	0	26	0	1	0	2
12	0	0	0	6	27	0	0	2	3
13	0	0	0	1	28	0	0	0	3
14	0	0	0	0	29	0	0	1	1
15	0	0	2	6	30	0	1	3	3
					Mean	0.13	0.20	0.76	2.30

Note. Logi = Logistic Regression Procedure.

Table 5: Results of LR (Number of Rejections Out of 100 Trials) for Sample Sizes 300 and 1000 under null conditions

	$d_\theta = 0$		$d_\theta = 1$			$d_\theta = 0$		$d_\theta = 1$	
Item	300	1000	300	1000	Item	300	1000	300	1000
1	7	7	7	8	16	6	8	8	6
2	2	8	4	6	17	4	6	2	2
3	4	8	8	3	18	4	2	4	5
4	11	4	7	8	19	6	4	5	3
5	2	7	6	4	20	6	4	6	4
6	6	10	8	3	21	4	8	7	4
7	10	7	7	3	22	1	4	9	5
8	4	6	7	7	23	11	8	3	7
9	6	3	6	7	24	3	5	5	7
10	5	3	1	8	25	6	4	3	9
11	3	4	2	5	26	4	2	4	7
12	3	8	8	6	27	9	10	7	10
13	6	6	5	4	28	7	5	1	5
14	7	7	3	3	29	7	7	9	7
15	5	4	6	5	30	3	6	4	2
					Mean	5.40	5.83	5.40	5.43

Note. LR = Likelihood Ratio Test.

Table 6: Results of DFIT (Number of Rejections Out of 100 Trials) for Sample Sizes 300 and 1000 under null conditions

	$d_\theta = 0$		$d_\theta = 1$			$d_\theta = 0$		$d_\theta = 1$	
Item	300	1000	300	1000	Item	300	1000	300	1000
1	3	7	6	9	16	7	10	16	9
2	5	10	4	10	17	5	9	7	7
3	8	7	5	4	18	4	3	12	12
4	6	5	7	8	19	2	3	4	2
5	8	8	7	2	20	6	5	2	2
6	7	10	6	5	21	4	8	8	8
7	5	3	4	0	22	7	1	0	0
8	4	6	6	7	23	3	5	0	7
9	5	2	5	6	24	1	2	4	2
10	6	0	1	2	25	8	2	0	1
11	7	6	8	10	26	8	0	4	5
12	5	6	5	5	27	6	6	3	8
13	8	3	3	4	28	0	3	0	0
14	4	6	9	6	29	3	4	6	7
15	3	6	4	2	30	2	4	3	0
					Mean	5.00	5.00	4.99	5.00

Note. DFIT = Differential Functioning of Items and Tests Procedure.

Table 7: Mean Type I error rates of Logi, LR, and DFIT (%)

	$d_\theta = 0$		$d_\theta = 1$	
	300	1000	300	1000
Logi	0.13	0.20	0.76	2.30
LR	5.40	5.83	5.40	5.43
DFIT	5.00	5.00	4.99	5.00

Note.

Logi = Logistic Regression Procedure;

LR = Likelihood Ratio Test;

DFIT = Differential Functioning Items and Tests Procedure.

Power study

Table 8 displays the Power study results of the three procedures. In its second column, we give the item numbers of the DIF items in the 7%, 10% and 20% DIF conditions. Table 9 and 10 displayed the mean true positive (TP) and false positive (FP) rates of the three procedures in the three manipulated factor conditions.

A true positive (TP) is an embedded DIF item which is correctly determined as DIF and an false positive (FP) is a non-DIF item which is falsely determined as DIF by the indices. TP rates are defined as the total number of the DIF items being detected across the 100 replications divided by the total number of the DIF items across the 100 replications. Similarly, FP rates are computed from dividing the total number of the non-DIF items being identified as DIF across the 100 replications by the total number of the non-DIF items across the 100 replications. For example, as shown in the 7% DIF, $d_\theta = 0$, and sample size of 300 condition of Table 8, items 4 and 6 were identified as DIF for respectively 1 and 90 times across the 100 trials. Because there were 2 DIF items in a single trial, the total number of DIF item across the 100 replications was 2*100 and hence the TP rate for the specified condition was $(1 + 90)/(2 * 100) = 0.455$.

The results in Table 9 indicate that as sample size increased, the mean TP rates increased for all procedures. For example, as sample size increased, the rates increased from 19% to 33% for Logi, a result consistent with French and Miller (1996). The increase for Logi, LR,and DFIT were about 14%, 16%, and 30%, respectively. DFIT, being as good as LR in Power, made the most improvement in large samples. Clearly, this can also be seen from

Table 8: Power Study Results (Number of Rejections out of 100 Trials) for Sample Sizes 300 and 1000

% of DIF	Item	$d_\theta = 0$						$d_\theta = 1$					
		300			1000			300			1000		
		Logi	LR	DFIT	Logi	LR	DFIT	Logi	LR	DFIT	Logi	LR	DFIT
7% DIF	4	1	82	20	20	100	97	6	79	60	40	100	100
	6	90	100	100	99	100	100	6	100	100	6	99	100
10%DIF	4	2	80	23	23	100	97	7	82	65	43	100	100
	6	95	100	100	96	100	100	6	100	100	5	100	100
	9	0	61	33	2	99	99	1	55	47	2	100	96
20%DIF	4	5	71	29	10	98	94	14	77	60	54	100	100
	6	72	100	100	92	99	100	7	100	100	2	100	100
	9	1	66	43	0	99	100	2	67	51	4	100	97
	11	0	100	100	11	99	100	2	99	99	3	100	100
	22	24	100	100	97	100	100	34	100	100	91	100	100
	30	0	33	25	1	83	76	0	25	20	1	67	63

Note. Logi = Logistic Regression Procedure;

LR = Likelihood Ratio Test;

DFIT = Differential Functioning of Items and Tests Procedure.

Table 9: Mean True Positive Rates (Power)

	Logi	LR	DFIT
Sample Size			
300	0.1922	0.8268	0.6677
1000	0.3342	0.9833	0.9737
Ability Distribution			
Equal	0.3830	0.9093	0.7840
Unequal	0.1433	0.9008	0.8573
%DIF			
7%	0.3350	0.9500	0.8463
10%	0.2350	0.8975	0.8003
20%	0.2195	0.8678	0.8155

Note. Logi = Logistic Regression Procedure;

LR = Likelihood Ratio Test;

DFIT = Differential Functioning of Items and Tests Procedure.

Table 10: Mean False Positive Rates

	Logi	LR	DFIT
Sample Size			
300	0.0055	0.0679	0.0566
1000	0.0151	0.0974	0.0795
Ability Distribution			
Equal	0.0019	0.0798	0.0699
Unequal	0.0187	0.0855	0.0662
%DIF			
0%	0.0085	0.0552	0.0499
7%	0.0112	0.0740	0.0633
10%	0.0095	0.0731	0.0624
20%	0.0120	0.1283	0.0966

Note. Logi = Logistic Regression Procedure;

LR = Likelihood Ratio Test;

DFIT = Differential Functioning of Items and Tests Procedure.

Table 8. The Power of all three procedures increased for nearly all the items as the sample size increased from 300 to 1000. In the 20% DIF and $d_\theta = 1$ condition, item 6, with a decrease from 7 to 2 in the number of correct identification for Logi, was the only exception. For item 22 in the 20% DIF condition, on the other hand, the Power of Logi made a huge progress when the sample size got larger. More specifically, the number of times item 22 being correctly identified of DIF increased from 24 to 97 and from 34 to 91 in the $d_\theta = 0$ and the $d_\theta = 1$ conditions, respectively.

As the difference in the mean ability between groups increased from zero to one, the detection rates decreased substantially for (about 24%) Logi, increased 7% for DFIT ,and were about the same for LR procedures. From Table 9 the detection rates decreased from 38% to 14% for Logi and increased from 78% to 86% for DFIT. The increase of DFIT mainly resulted from item 4 in the sample of size 300 where its number of correct identification increased from 20 to 60, 23 to 65, and 29 to 60 respectively in the 7%, 10%, and 20% DIF conditions. And the decrease of Logi was mainly from item 6 in all DIF percentage conditions. For example, in sample size of 300, as d_θ changed from 0 to 1, the number of correct identification out of the 100 trials decreased from 90 to 6, 95 to 6, and 72 to 7 respectively in the 7%, 10%, and 20% DIF conditions.

In summary, among the three DIF percentage levels, the mean Power rates in the 7% level were the highest for all the three procedures. The mean Power rates for DFIT did not differ much whether the tests had 10% or 20% DIF items. For LR, there was an 3% increase in the mean Power for the 10% level over the 20% level, and for Logi, the decrease was about

1.35%.

From the mean FP rates displayed on Table 10, we found under each level of the three manipulated factor conditions, the procedure with higher mean TP rates also had higher FP rates. For instance, in samples of size 300, the orders of TP rates and FP rates, from the highest to the lowest, were both LR, DFIT, and Logi. As sample size got larger, the FP rates increased for all procedures. LR had the highest TP rates (up to 98%) and FP rates (up to about 9.7%) in both sample sizes, while Logi had the smallest TP rates and FP rates. And when d_θ changed from 0 to 1, the FP rates increased less than 2% for Logi and LR, and were about the same for DFIT (from 6.9% to 6.6%). For tests containing 20% DIF items, the FP values were much too high for LR (12.8%) and DFIT (9.7%). In other percentages of DIF conditions, the FP rates were acceptable with a range from 0.2% to 7.4%.

Real Data Analysis

In addition to comparing the performances of the three procedures in the above simulation study, we would like to investigate how their DIF results would differ in practice. Next, we applied the three procedures to analyze a real dataset.

Cognitive Abilities Screening Instrument (CASI)

The Cognitive Ability Screening Instrument (CASI) is suitable for clinical use in screening dementia for its simplicity and little time required. Its effectiveness in the screening of dementia has been recognized cross-culturally. The CASI sum score, ranging from 0 to

100, is a linear combination of the item scores, and is used to represent the underlying cognitive ability of the patient. The validity of the Chinese version 2.0 (CASI C-2.0) with 45 polytomously scored items has been well studied. Lin et al. (2002) suggest the use of different cutoff scores for different education and age groups while screening dementia. If the CASI sum score is taken as the measure of the patient's underlying cognitive ability, a patient is classified into the group of questionable dementia if his or her cognitive ability, i.e., the CASI score, falls below a cutoff threshold determined from past research. However, the use of different cutoff scores for different demographic groups seems to imply that not the cognitive ability alone contributes to a higher score on the test. In other words, for patients with the same amount of cognitive abilities, they perform differently on some items as well as on the whole test, depending on their demographic background such as the gender of the patients and so on. The phenomena strongly suggest the existence of potential DIF items in CASI C-2.0.

Recently Crane et al. (2004), in a survey study of 2940 individuals, presented a DIF analysis of CASI using ordinal logistic regression procedure (Zumbo, 1999) under both the IRT scoring (using PCM) and the traditional CASI scoring. They found a disappointingly large number of DIF items with respect to at least one of the demographic variables. To investigate the problem of DIF in the Chinese version, here we conduct a DIF analysis of the 45 items with the Logi, LR, and DFIT procedures with respect to age, gender, and education groups under the GRM. Our results are then compared to that of Crane et al. (2004).

Subjects and Methods

Data from 608 patients (289 males, 319 females) visiting the Memory Clinic of the Neurological Institute of the Taipei Veterans General Hospital (Taipei VGH) were used in this analysis. Their ages ranged from 34 to 95 years (Mean=73.57, SD=8.10) with 78% and 22% respectively in groups of below and no less than 80 years old. Their education period of time ranged from 0 to 25 years (Mean=8.32, SD=5.68). If we divide them into three education levels : (a) no schooling (Edu=0) , (b) elementary-school education (Edu=1-6), and (c) beyond (Edu$\geq$7), there would be 105 (17.3%), 164 (27.0%), and 339 (55.7%) subjects within each level respectively.

For each procedure, we consider the group assignment with respect to their age (age=2 for subjects no less than the Age of 80 and Age=1 otherwise), gender, and education level (Edu=1 for subjects with no schooling and Edu=2 otherwise). In Logi, the Bonferroni-adjusted p-value is adopted. That is, nonuniform DIF is said to occur when the original p-value multiplied by 45 is less than α of the value 0.05. The critical values for the statistics G^2 in LR are 5.991, 7.815, 9.488, 11.07, and 12.59 for items with the numbers of categories as 2, 3, 4, 5, and 6 respectively. In DFIT, we adopted 0.088, the critical value of NCDIF for the sample size of 300 and $d_\theta = 1$ condition in our simulation study, to be the critical value in CASI for their similar features in sample size and the potential mean difference in the groups.

DIF Results

The DIF results of the CASI items using Logi, LR, and DFIT procedures are showed in Tables 11 and 12, together with the results of Crane et el. (2004) using IRT scoring, where a check mark "v" is placed in the appropriate cell to indicate significant DIF. In the three procedures, there were respectively 8, 17, 28 items detected as DIF with respect to at least one of these demographic variables whereas the analysis of Crane et al. (2004) showed 10 items exhibiting DIF.

The items detected as DIF with respect to at least one of the demographic variables by all the three procedures were AGE, DBB, BODY, OBJA, OBJB, and RCOJ (see appendix for descriptions of the item abbreviations). Among the six items, AGE, BODY, and RCOJ were detected as DIF with respect to both the Gender and Edu status, while OBJA to the Edu status. There is no overlap between the results in Crane et al. (2004) and those detected here using Logi.

Among the three demographic variables, the number of items detected as DIF with respect to the Age status is the smallest. More specifically, there were respectively 0, 1, and 10 items in Logi, LR, and DFIT for Age status. The amount of DIF items with respect to the Gender and Edu status were, on the other hand, about the same respectively in Logi and in LR. In DFIT, there were 16 and 23 items detected as DIF to Gender and Edu status, respectively. Moreover, the items BODY and RCOJ were detected as DIF most frequently across all procedures, each for 8 and 7 times out of 12 trials respectively.

With such a large discrepancy between their detection results, a thorough DIF study of

Table 11: Evaluation of DIF for Age, Gender, and Education in CASI

	Crane			Logi						LR			DFIT			
Item	Age	Gender	Edu	Age		Gender		Edu		Age	Gender	Edu	Age	Gender	Edu	Total
				NU	U	NU	U	NU	U							
AGE							v		v		v	v		v	v	6
MONTH																
NEWYR											v			v		2
MNT															v	1
SUN											v					1
MOON											v					1
RGS1	v										v			v		2
YR																
MO									v						v	2
DATE								v	v				v		v	4
DAY																
TDAY												v				1
ANML											v	v	v	v	v	5
RC1A																
RC1B															v	1
RC1C															v	1
SPB																
SPA														v		1
DBA															v	1
DBB	v					v						v			v	3
DBC												v				1
SUBA																
SUBB											v					1
SUBC													v			1
SIM1			v													
SIM2			v										v	v	v	3
SIMA			v										v	v	v	3

Table 12: Evaluation of DIF for Age, Gender, and Education in CASI (Conti.)

	Crane			Logi						LR			DFIT			
Item	Age	Gender	Edu	Age		Gender		Edu		Age	Gender	Edu	Age	Gender	Edu	Total
				NU	U	NU	U	NU	U							
JGTA																
JGTB															v	1
JGTC																
RPTA	v															
RPTB															v	1
READ														v	v	2
DRAW1			v										v	v	v	3
DRAW2			v											v	v	2
DRAW3			v									v		v		2
WRT												v	v	v	v	4
CMD											v	v			v	3
RC2A		v											v		v	2
RC2B																
RC2C	v															
BODY							v		v	v	v	v	v	v	v	8
OBJA								v			v	v		v	v	5
OBJB							v					v		v	v	4
RCOJ							v		v		v	v	v	v	v	7
Total	4	1	6	0	0	1	5	1	4	1	11	12	10	16	23	
				0		6		5								
	11					8					17			28		

Note. Crane = Crane et al.;

Logi = Logistic Regression Procedure;

LR = Likelihood Ratio Test;

DFIT = Differential Functioning of Items and Tests Procedure.

the items in CASI by experts from other aspects such as test construction is much needed. Without the knowledge of which items are known to exhibit DIF, we were unable to compare the performance of the three procedures in accurately judging biased items. Nevertheless, according to the above simulation studies, LR is shown to be more reliable with greater power and reasonable Type I error rate. Using LR, we therefore found about 17 out of 45 items exhibiting DIF in CASI C-2.0. In other words, there is a large percentage of DIF items in CASI C-2.0.

One of the limitations of this analysis is the lack of sufficient number of examinees for obtaining stable IRT parameter estimates, especially for the group of Edu=1 which only consisted of 105 patients. In addition, the majority of the samples in our study were patients with Alzheimer's disease and those in Crane's study, on the other hand, were from a survey in a cohort study of the elderly. More representative samples in Taiwan would be needed to make a cross-cultural comparison of the DIF analysis. Besides, the criterion we adopted for NCDIF in DFIT were generated from data of 4-category items whereas the numbers of categories of CASI items range from 2 to 6. A better critical value for NCDIF would be necessary for a more reliable DIF analysis using DFIT.

Discussion and Conclusion

In most cases, the detection rates of the three procedures were higher with larger sample size, smaller d_θ, and less percentage of DIF items contained in a test. DFIT, however, performs better with unequal ability distribution as found by Flowers et al. (1999), and its mean TP

rate in the 20% DIF condition was slightly higher than that in the 10% DIF condition.

LR performs best in Power under all conditions. The performance of DFIT was not as good as LR, but was also quite acceptable. The Power and Type I error of DFIT would not be much affected by the percentage (above 10%) of DIF items contained in a test, hence for data of which unequal ability distributions and large percentage of DIF items are expected, DFIT is recommended. However, under a condition with no knowledge of ability distributions and percentage of DIF items in a test, we would suggest the use of LR. In addition, as mentioned above that the comparison of DFIT to other DIF indices might be ambiguous because the good performance of DFIT might be an artifact resulting from using the empirical NCDIF critical values. In analyzing real data, finding an empirical critical values for DFIT might be a challenging and time-consuming task which requires more investigation in the future.

The performance of Logi in Type I error and Power are both much lower than those of LR and DFIT. The extremely low Power performance of the continuation ratio logits seemed to be quite different from the results of French and Miller (1996), though the PCM model was used instead in their study. To search the possible cause, we inspected its performance on all the items in Table 8 and found that all but items 6 and 22 under some conditions yielded poor TP rates. For example, Logi worked for items 6 only under the $d_\theta = 0$ condition and when $d_\theta = 1$, the numbers of correct identification of DIF were all below 10 out of 100 trials for item 6. After examining the item parameters of the two items, 6 and 22, we found that the differences of their item parameters between the reference and focal groups were the largest among all DIF items in the study (see Table 1 and Table 2). Besides, most of

the DIF items used in French and Miller (1996) contained larger DIF sizes than what we had in our study. In our opinion, one possible explanation is that Logi was only capable of detecting DIF with relatively large DIF size. However, future studies are needed to confirm such a speculation. In addition, Narayanan and Swaminathan (1996) have shown that the nonuniform detection rates for Logi in dichotomous items are affected by the type of item. The order in detection rates from the highest to the lowest was (1) low b, high a; (2) median b, high a; (3) high b, low a; and (4) median b, low a for a two-parameter IRT model. It would not be of much surprise to see that item type still affects the performance of Logi on detecting DIF of the polytomous scored items as well.

Apparently, this study showed inefficiency of Logi while compared to the other methods. It is quite disappointing because both LR and DFIT have their shortcomings. For LR, the calibration of the 31 models are required in each simulation trial for a thirty-item test using MULTILOG. With 100 replications for each condition, it is a time-consuming job. In addition, the NCDIF critical values of DFIT have not been well established whereas Logi has the distinct advantages of telling whether an item exhibits uniform or nonuniform DIF.

There are four major limitations in this study. One is that we only compare the three procedures under the graded response model framework. Other polytomous IRT models which have been less studied for DIF in the literature, such as the partial credit models within the Rasch family, have gained popularity in other applied fields. Thus, it would become practically useful to extend the comparison of the different DIF indices under other types of polytomous IRT models.

Secondly, the indices used for Logi here were proposed by Maldonado and Greenland (1993). Other indices, such as the effect size measure (Jodoin, 2001), could also be investigated.

Thirdly, in most DIF studies, one single DIF item is commonly used to avoid the contamination of multiple DIF items in a test. However, we found such scenario to be of theoretical interest only since in analyzing real data, more than one items are usually identified to have DIF and the benchmark of a variety of indices established under simulations with one single DIF item might not be appropriately applicable. However, as the number of DIF items increases and the effect of contamination among the items unknown, we are unable to assess how much the results depend on the particular choice or combination of item parameters, or their interaction with the number of categories and so on. Especially with the extremely different performance of Logi on the various items, it is impossible for us to locate the causes of such results. Moreover, in the simulation study, the items are simultaneously instead of sequentially judged as DIF within each detection. In the sequential process, one and only one item is said as DIF and consequently taken out in each detection until no better performance can be achieved from taking more DIF items. The sequential technique has the advantage that the test structure would be adjusted closer and closer to the real structure in the successive removal of DIF items. But sequential technique requires tremendous effort and is therefore difficult to carry out, especially in a simulation study, and therefore we simply adopt the simultaneous technique. To what extent the use of simultaneous or sequential detection affects the results of the simulation study is unknown.

Lastly, as in most simulation studies in the literature where the same number of categories are used for all the items in a test, we chose four categories for each item throughout the study. However, to what extent the number of categories affects the Power of the DIF indices is yet unknown. For example, for the items in CASI, their numbers of categories range from 2 to 6 and the choice between using a general critical value or different critical values for items with different number of categories for the DFIT index requires more investigation.