A Comparison of Three Polytomous DIF Detection Methods

Li-An Wu* Rung-Ching Tsai**

*Department of Statistics,
National Chiao-Tung University

**Department of Mathematics,
National Taiwan Normal University

Abstract

The performance of the three procedures -- the logistic regression procedure (LogR), the likelihood ratio test (LRT), and the differential functioning of items and tests procedure (DFIT) in detecting differential item functioning (DIF) under the graded response model were compared in a simulation study. Factors manipulated included sample size, differences in the ability distributions between the focal and the reference groups, and four different percentages of DIF items contained in a test. For each of the sixteen combinations, 100 replications of DIF detection were simulated. All three DIF procedures adhered to nominal type I error rates under most conditions. LRT was the most powerful among the three under all situations. DFIT was less powerful than LRT, but also useful for DIF detection especially with groups of different ability distributions and relatively large percentage of DIF items. LogR, with mean powers lower than 0.4 in all conditions, appeared to be sensitive only to items with large DIF size.

Keywords: differential item functioning (DIF), logistic regression procedure, differential functioning of items and tests procedure.

三種多元化計分題之試題差異性診斷法的比較

吳莉安[1] 蔡蓉青[2]

[1]國立交通大學統計研究所

[2]國立台灣師範大學數學系

摘要

本論文以模擬研究比較了三種不同的試題差異性（DIF）診斷法—羅吉斯迴歸檢定、概度比檢定，以及差異試題及測驗功能檢定在等級反應模式（graded response model）下之表現。操縱變因包括了樣本數（兩種）、母群體之分配（兩種）、以及測驗中所含DIF題數之比例（四種）。在十六種組合之下，各做了一百次試驗。試驗結果發現，這三種方法之型一誤差（type I error）大致上都符合0.05的限定。而在檢定力（power）的表現上，概度比檢定最好、差異試題及測驗功能檢定次之、羅吉斯迴歸檢定最差。平均而言，羅吉斯迴歸檢定之檢定力的表現低於0.4，而且只對DIF性質明顯的題目偵測較為靈敏。

關鍵字：試題差異性、羅吉斯迴歸檢定、差異試題及測驗功能檢定

I、Introduction

Differential item functioning (DIF) is said to be present when examinees from two different groups with the same 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), Penfield and Lam (2000), Teresi and Fleishman (2007) and Mapuranga, Dorans, and Middleton (2008). Potenza and Dorans (1995) and Mapuranga et al. (2008) provide classifications of the existent DIF methods according to a variety of different criteria. For instance, the DIF procedures were classified using the DIF, statistical and practical criteria in Mapuranga et al. (2008).

Here we distinguish between parametric and nonparametric DIF methods. Parametric DIF procedures could be roughly 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 (Load, 1980). The second one is to compare areas between the item characteristic curves or the expected item score functions estimated from the two groups, such as the differential functioning of items and tests procedure (DFIT; Cohen, Kim, & Baker, 1993; Flowers, Oshima, & Raju, 1999; Kim & Cohen, 1991; Oshima, McGinty, & Flowers, 1994; Raju, van der Linden, & Fleer, 1995). The last class is to compare the likelihood functions, as in the likelihood ratio test (LRT; Thissen, Steinberg, & Gerard, 1986) and the logistic regression procedure (LogR; Swaminathan & Rogers, 1990; French & Miller, 1996). Furthermore, the multiple indicator, multiple causes (MIMIC) confirmatory factor analysis model within the structural equation modeling framework has also gradually gained its popularity as an approach to DIF analysis (Muthén, 2002). Nonparametric DIF procedures which do not assume a specific statistical model, such as the Mantel-Haenszel procedure (MH; Mantel & Haenszel, 1959; Holland & Thayer, 1988; Zwick, Donoghnue, & Grima, 1993), the simultaneous item bias procedure (SIBTEST; Shealy & Stout, 1993) and its polytomous extension (Poly-SIBTEST; Chang, Mazzeo, & Roussos, 1996) are also commonly used.

LogR 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 LogR 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. Raju et al. (1995) presented DFIT and offered an empirical demonstration using dichotomous data. It was extended to the dichotomous multidimensional case by Oshima, Raju, and Flowers (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 (French & Miller, 1996; Raju et al., 1995; Oshima et al., 1997; Flowers et al., 1999).

In a comparison to MH (Holland & Thayer,1988) and SIBTEST (Shealy & Stout, 1993) for dichotomous items, LogR was found as effective as MH and SIBTEST in detecting uniform DIF (Narayanan & Swaminathan, 1994; Rogers & Swaminathan, 1993). In the detection of nonuniform DIF, LogR was shown more powerful than MH and as powerful as SIBTEST (Narayanan & Swaminathan, 1996). Bolt (2002) compared

the performance of LRT, DFIT and Poly-SIBTEST procedures under graded response models (GRM; Samejima, 1969) in terms of their type I error and power. When data were generated from GRM, LRT was superior to DFIT in its type I error and power. LRT and DFIT were both better than Poly-SIBTEST with small sample sizes. The majority of DIF studies in the literature have been done using GRM, partly because softwares are readily available for its estimation and testing.

LogR 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. Despite many studies have been conducted to compare the performance of different DIF procedures using simulated data sets, the performance of LogR has not yet been compared to both LRT and DFIT in the literature. To evaluate LogR in comparison to LRT and DFIT, we investigate their performances under GRM through a simulation study.

Paralleled with Bolt's (2002) study in a number of design factors such as the choice of DIF indices of LRT and DFIT and the use of GRM as the IRT model fitted to the polytomous data, our study compared the performance of LogR instead of Poly-SIBTEST due to the growing interest in using LogR for DIF analysis. However, as pointed out by one of the reviewers that our use of logistic regression in LogR differed from those in the literature (Rogers & Swaminathan, 1993; French & Miller, 1996) in that we actually used the ability estimates under GRM rather than the sum score in the regression. By doing so, we made the comparison among the three DIF indices under consideration more of a parametric nature while the likelihood, the response function and the parameter estimates were all obtained from GRM. To what extent does this change affect the effectiveness of LogR is yet unknown and perhaps worth further investigation. Similar strategy of using ability estimates in LogR has in fact been adopted in Crane, Belle, and Larson (2004).

Moreover, one of the aims of Bolt's (2002) study was to investigate the relative importance in selecting the "correct" model in evaluating the effectiveness of the DIF indices and therefore generalized partial credit model (GPCM; Muraki, 1992) or the two-parameter sequential response model (2p–SRM; Mellenbergh, 1995) were used for data simulation in their study. In contrast, we were interested in the effectiveness of the DIF indices under different levels of item contamination and therefore different percentages of items containing DIF were also manipulated to see the performance of those DIF indices in real tests, which was different from Bolt's design with a fixed number of items (i.e., 30 items + 1 DIF item).

A、Graded-response model (GRM)

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 category k or above to item i, with possible scoring $k = 1,2,\cdots,m$, is:

$$P_{ik}^{*}(\theta_l) = \frac{exp[a_i(\theta_l - b_{ik})]}{1 + exp[a_i(\theta_l - b_{ik})]}, \quad k = 2,\cdots,m \tag{1}$$

where a_i is the discrimination parameter for item *i*, b_{ik} is the category threshold parameter associated with category *k* of item *i*, and $b_{ik'} > b_{ik}$ whenever $k' > k$, 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 (Samejima, 1969). The probability of scoring in each category is computed from adjacent boundary probabilities. For example, the probability of examinee l responding category *k* to item *i*, is

$$P_{ik}(\theta_l) = P^*_{ik}(\theta_l) - P^*_{i,k+1}(\theta_l), \quad k = 1,2,\cdots,m \tag{2}$$

with $P^*_{i,1}(\theta_l) = 1$ and $P^*_{i,m+1}(\theta_l) = 0$. Equation (2) is referred to as the item category response function (ICRF).

In the following, we give details of the three DIF indices under consideration here, followed by the design and results of the simulation studies.

B、DIF indices

Method 1: Logistic regression procedure (LogR)

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 membership. 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 continuation-ratio logit model used was:

$$\log it(\rho_{ik} \mid \theta_l, g) = \log\left(\frac{\rho_{ik}}{1-\rho_{ik}} \mid \theta_l, g\right) = \alpha_{ik} + \beta_{i1}\cdot\theta_l + \beta_{i2}\cdot g + \beta_{i3}\cdot(g\cdot\theta_l) \tag{3}$$

where the left side of the function is the continuation ratio logit, ρ_{ik} is the probability of response *k* on item *i* , given that the examinee's response is category *k* or higher, i.e.,

$$\rho_{ik}(\theta_l) = \frac{P_{ik}(\theta_l)}{P_{ik}(\theta_l) + \cdots + P_{im}(\theta_l)}, k = 1,\cdots,m-1.$$

θ_l is the ability level of examinee l, and g represents his or her group membership, either as the reference or the focal group. Note that the above use of logistic regression in LogR differed from those in the literature (Rogers & Swaminathan, 1993; French & Miller, 1996) in that here we used the θ_l's under GRM rather than the sum score as the person's ability estimate. Simulation results suggested that different nested models were better examined separately for detecting uniform and nonuniform DIF items (Jodoin & Gierl, 2001). Thus, we first 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 the significance level α, nonuniform DIF was detected for item i. If not, the interaction term would be eliminated for the following analysis.

Maldonado and Greenland (1993) investigated a variety of confounder-selection strategies in deciding whether the covariate should be included to adjust for the exposure effect in Poisson regression models and found that both the "change-in-estimate by an amount of 10%" and "equivalence-test-of-the-difference" strategies performed best among all criteria in a simulation study. We adopted their approach in determining uniform DIF items such that for items free of nonuniform DIF, we then compare the β_1 estimates of the models with and without grouping variable g. If a 10% difference between the β_1 estimates with and without grouping variable was found in the model, uniform DIF was said to be present for this particular item (Crane et al, 2004).

Method 2: Likelihood ratio test (LRT)

Suppose the null and alternative hypotheses for the model parameter(s) γ are respectively

H0 : $\gamma \in Model_c$ (which contains *N* free parameters), and

HA : $\gamma \in Model_A$ (which contains *N*+*K* free parameters).

The likelihood ratio (LR) of interest for these two models is defined as

$$LR = \frac{L^*(Model_c)}{L^*(Model_A)}$$

where $L^*(Model)$ refers to the maximum of the likelihood function $L(Model)$ under the specified model. Define G^2 as twice the negative natural log transformation of the likelihood ratio of the two models such that

$$G^2 = -2\ln(LR) = -2\ln(L^*(Model_c)) - [-2\ln(L^*(Model_A))]$$

In large samples, G^2 is approximately distributed as a chi-square variate with *k* degrees of freedom.

While using the likelihood ratio test in a DIF analysis under GRM (Kim & Cohen, 1998), the compact model ($model_c$) of the null hypothesis is defined by constraining all item parameters to be equal for the reference and the focal groups, whereas in the augmented model ($model_A$) of the alternative hypothesis where only the parameters of the studied item are allowed to differ across groups. If the test statistic G^2 exceeds a critical chi-square (χ^2) value at a given level of α , the null hypothesis of no DIF is rejected for that particular item.

The type I error rates of LRT have been found consistent with the nominal α level when data were generated from GRM (Kim & Cohen,1998). Also when LRT was applied to polytomous data, the GRM has been a commonly chosen model in previous studies (Ankenmann, Witt, & Dunbar, 1999; Reise, Widaman, & Pugh, 1993).

Method 3: Differential functioning of items and tests procedure (DFIT)

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,

$$ES_{iR}(\theta_l) \neq ES_{iF}(\theta_l)$$

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

$$ES_{ig}(\theta_l) = \sum_{k=1}^{m} kP_{ik,g}(\theta_l),$$

where *g*=*R* or F. The difference between the two expected scores, defined as $d_i(\theta_l) = ES_{iF}(\theta_l) - ES_{iR}(\theta_l)$, 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 and facilitates the definition of CDIF for each item as

$$CDIF_i = Cov(d_i, D) + \mu_{d_i}\mu_D \tag{4}$$

where μ_{d_i}, μ_D are the means of *di* and $D(=\sum_{i=1}^{n} d_i)$, respectively. Under the assumption that all but the studied item are free from DIF, $d_j = 0$ for all $j \neq i$ (i.e., $D = d_i$), where *i* is the item studied, another index, the noncompensatory DIF (NCDIF), is defined as

$$NCDIF_i = \sigma_{d_i}^2 + \mu_{d_i}^2,$$

where $\sigma_{d_i}^2$ is the variance of *d*. Or equivalently, for each item *i*,

$$NCDIF_i = E_F(d_i(\theta)^2) = \int_{\theta} d_i(\theta)^2 f_F(\theta) d\theta \tag{5}$$

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 chi-square distribution under the null hypothesis of no DIF. However, the chi-square 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.

II、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. Most item parameters were taken from those in Bolt's study, except that five response categories were used in Bolt's study and four in ours. Among

the biased items, items 4, 6, 22 and 30 were with uniform DIF whereas items 9 and 11 were with nonuniform DIF.

A、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.

First, two sample sizes were considered. Responses from 300 examinees were generated for each group as the small size for its sufficiency in obtaining stable IRT item parameter estimates. The other sample size of 1000 for each group was used.

Secondly, two different ability distributions were assumed for the focal group. In the "Equivalent" condition (i.e., $d_\theta = 0$), both the focal and reference groups had the ability distributions of a standard normal N(0,1). In the "Nonequivalent" condition, the focal group was sampled from an N(-1,1) distribution whereas the reference group was sampled from the N(0,1) distribution. That is, there existed a difference of 1 in the mean ability between the focal and reference groups, i.e., $d_\theta = 1$ in the non-equivalent condition.

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 was replicated 100 times to facilitate type I error and power calculations of the three DIF detection procedures.

B、Estimation and criteria

In the logistic regression procedure (LogR), the estimates of examinees' ability θ_l'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, the number of items in the test, was less than $\alpha = 0.05$, the item is said to contain nonuniform DIF.

For items lack of nonuniform DIF, we examined whether they exhibited uniform DIF by comparing the β_1 estimates with and without grouping variable in the models as described above. An 10% difference in the coefficient estimates indicated the existence of uniform DIF (Crane et al, 2004). The logistic regression models for logR were done using S-PLUS.

In the LRT test, DIF of each item is studied by comparing "negative twice the loglikelihood" 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 in the numbers of item parameter between the augmented model and the compact model equals to four. Hence, the likelihood ratio statistic G^2 is approximately distributed as a

χ^2 distribution with four degrees of freedom under the hypothesis of DIF free. So if G^2 value exceeds 9.488, the critical value for $\chi^2(4)$ at $\alpha = 0.05$, the item is judged as exhibiting DIF.

In using DFIT in the DIF analysis, we first estimated item parameters of each group separately by MULTILOG (Thissen, 2003). EQUATE (Baker, 1993) which implemented the test characteristic curve-based Stocking and Lord procedure for equating (Stocking & Lord, 1983) 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 (5),

$$NCDIF_i = \int_\theta d_i(\theta)^2 f_F(\theta) d\theta = \int_\theta \frac{1}{\sqrt{2\pi}} d_i^2(\theta) \exp[-\frac{\theta^2}{2}] d\theta ,$$

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 were determined by the distributions of NCDIF values for data obtained from the tests containing no DIF items (Bolt, 2002). In this study, the chosen critical values correspond to an α value of 0.05 in the distribution of the $30 \cdot 100 = 3000$ NCDIF values obtained from the four null conditions. For 300 per group, the critical values for the equivalent and non-equivalent conditions were respectively 0.134 and 0.088 whereas for 1000 per group, their critical values were 0.039 and 0.026.

III、Results

A、Type I error study

Tables 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 in 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, in Table 4 the results of LogR 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 significance level 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 the LogR results in Table 4, the number of rejections was higher in the $d_\theta = 1$ condition than in the $d_\theta = 0$ condition for nearly every item for both sample sizes. For example, we found in the samples 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 for the samples of size 1000. For each of the d_θ levels, 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 LogR would increase. However, with respect to the nominal type I error rate 5%, the type I error of LogR was relatively low with the maximum of 2.3%.

From the LRT results in Table 5, we found that in the samples of size 300, the type I error rates for $d_\theta = 0$ and $d_\theta = 1$ were both 5.40%. But for those of size 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 LogR, only half of the items had greater number of rejections in the $d_\theta = 1$ condition than in the $d_\theta = 0$ condition. 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 LRT, 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, LogR has the lowest rates. The rates of LRT 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% for LogR, 5.83% for LRT, and 5.00% for DFIT. In comparison to our results, the mean type I errors of LRT under GRM in Bolt's study had a larger range from 4.9% to 6% and a slightly smaller average of 5.3% for the four conditions.

B、Power study

Table 8 displays the results of the power study 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 display 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 $2 \cdot 100 (= 200)$ and hence the TP rate for the specified condition was $\frac{(1+90)}{(2 \cdot 100)} = 0.455$.

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 LogR, a result consistent with French and Miller (1996). The increase for LogR, LRT,and DFIT were about 14%, 16%, and 30%, respectively. DFIT, being as good as LRT in its power, made the most improvement in large samples. Clearly, this can also be seen from 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 LogR, was the only exception. For item 22 in the 20% DIF condition, on the other hand, the power of LogR made a huge progress when the sample size got larger. More specifically, the number of times item 22 being correctly identified as DIF increased from 24 to 97 and from 34 to 91 in the $d_\theta = 0$ and $d_\theta = 1$ conditions, respectively.

As the difference in the mean ability between groups increased from zero to one, the detection rates decreased substantially for (about 24%) LogR, increased 7% for DFIT ,and were about the same for LRT procedures. From Table 9 the detection rates decreased from 38% to 14% for LogR 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 LogR 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 LRT, there was an 3% increase in the mean power for the 10% level over the 20% level, and for LogR, the increase was about 1.55%. The decrease in power from the 10% to 20% DIF conditions was possibly caused by the fact that item 9 was added to the power calculation. To be more specific, we can see from Table 8 that in most cases the rejections of item 9 were much lower than items 4 and 6 under the same condition. Thus the mean power (total rejections out of all the biased item trials) is smaller while item 9 was added in the averaging. Similarly, of the three biased items 11, 22 and 30 added, only item 22 exihibiting larger DIF size and thus the mean power decreased.

The design with 31 studied item in Bolt's study, is closest to our 7% DIF item study, among all the contamination rates under consideration. Their mean power of LRT, 0.93, is close to 0.95 of ours whereas the mean power of DFIT as 0.71 appears to be much lower than 0.84 obtained in our results. Moreover, as the sample size increases, the power also increases. Under the same sample size, the power of LRT is larger than that of DFIT.

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 LRT, DFIT, and LogR. As sample size got

larger, the FP rates increased for all procedures. LRT had the highest TP rates (up to 98%) and FP rates (up to about 9.7%) in both sample sizes, while LogR 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 LogR and LRT, and were about the same for DFIT (from 6.9% to 6.6%). In addition, we found across different levels of DIF percentage that items with larger type I errors also have larger FP rates. For tests containing 20% DIF items, the FP values were much too high for LRT (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%. The FP rates of LogR under all conditions were too low (less than 2%).

IV、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.

LRT performs best in power under all conditions. The performance of DFIT was not as good as LRT, 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 the condition with no knowledge of the ability distributions and percentage of DIF items in a test, we would suggest the use of LRT. 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 empirical critical values for DFIT might be a challenging and time-consuming task which requires more investigation in the future.

The performance of LogR in type I error and power are both much lower than those of LRT 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. Note that we used the ability estimates under GRM rather than the sum score in our logistic regression for LogR. 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, LogR worked for item 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 (up to 1) among all DIF items in the study (see Table 1 and Table 2). The parameter differences of items 4, 9 and 11 were the second largest (about 0.5). While comparing the powers of all the biased items under the 20% DIF condition, we found that LogR appears more powerful in detecting uniform DIF (item 4) than non-uniform DIF (items 9 and 11). Moreover, for uniform DIF items like items 4 and 22, the powers of LogR highly depend on the DIF sizes of the biased items.

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 LogR was only capable of detecting items with relatively large DIF sizes. Another reason could be that the "change in β_1 estimates by an amount of 10%" criterion adopted here for identifying uniform DIF for LogR might be too strict and the 10% difference in paramter estimates could be lowered for better detection of uniform DIF. However, future studies are needed to confirm such speculations. In addition, Narayanan and Swaminathan (1996) have shown that the nonuniform detection rates for LogR 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) medium b, high a; (3) high b, low a; and (4) medium 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 LogR on detecting DIF of the polytomous scored items as well.

Apparently, this study showed inefficiency of LogR while compared to the other methods. It is quite disappointing because both LRT and DFIT have their shortcomings. For LRT, 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 LogR 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 LogR here were proposed by Maldonado and Greenland (1993). Other indices, such as the effect size measure (Jodoin & Gierl, 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. Thus, the benchmark of a variety of indices established under simulations with one single DIF item might not be appropriately applicable. 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 LogR on the various items, it is impossible for us to identify 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 more effort and is therefore labor intensive 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.

V、Acknowledgement

This research was partly supported by the grant from the National Science Council of Taiwan (NSC 92-2413-H-003-068).

References

Ankenmann, R. D., Witt, E.A., & Dunbar, S.B. (1999). An investigation of the power of the likelihood ratio goodness-of-fit statistic in detecting differential item functioning. *Journal of Educational Measurement, 36*, 277-300.

Baker, F. B. (1993). EQUATE2: Computer program for equating two metrics in item response theory [Computer program]. Madison: University of Wisconsin, Laboratory of Experimental Design.

Bock, R. D., & Aitkin, M. (1981). Maximum likelihood estimation of item parameters: an application of the EM algorithm. *Psychometrika, 46*, 443-459.

Bolt, D. M. (2002). A monte carlo comparison of parametric and nonparametric polytomous dif detection methods. *Applied Measurement in Education, 15*, 113-141.

Camilli, G., & Shepard, L. A. (1994). *Methods for Identifying Biased Test Items*. Sage: Thousand Oaks.

Chang, H. H. & Mazzeo, J. (1994). The unique correspondence of the item response function and item category response functions in polytomously scored item response models. *Psychometrika, 59*, 391-404.

Chang, H. H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 32*, 79-96.

Cohen, A. S., Kim, S. H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement, 17(4)*, 335-350.

Crane, P. K., Belle, G. V., & Larson, E. B. (2004). Test bias in a cognitive test: differential item functioning in the CASI. *Statistics in Medicine, 23*, 241-256.

Flowers, C. P., Oshima, T. C., & Raju, N. S. (1999). A description and demonstration of the polytomous-DFIT framework. *Applied Psychological Measurement, 23*, 309-326.

French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement, 33*, 315-332.

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: Lawrence Erlbaum.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*, 329-349.

Kim, S. H., & Cohen, A. S. (1991). A comparison of two area measures for detecting differential item functioning. *Applied Psychological Measurement, 15(3)*, 269-278.

Kim, S. H., & Cohen, A. S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement, 22*, 345-355.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale NJ: Erlbaum.

Maldonado, G., & Greenland, S. (1993). Simulation study of confounder-selection strategies. *American Journal of Epidemiology, 138*, 923-936.

Mantel, N., & Haenszel, W.M. (1959). Statistical aspects of the analysis of data from retrospective studies of diserse. *Journal of National Cancer Institute*, 22, 719-748.

Mapuranga, R., Dorans, N. J., & Middleton, K. (2008). A review of recent developments in differential item functioning, Paper presented at the annual meeting of the National Council on Measurement in Education (NCME) held in March, 2008, New York.

Mellenbergh, G. J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement, 19*, 91-100.

Miller, T. R., & Spray, J. A.(1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement, 30*, 107-122.

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

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika*, *29*, 81-117.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement, 18*, 315-338.

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

Oshima, T. C., McGinty, D., & Flowers, C. P. (1994). Differential item functioning for a test with a cutoff score: use of limited closed-interval measures. *Applied Measurement in Education, 7(3)*, 195-209.

Oshima, T. C., Raju, N. S., & Flowers, C.P. (1997). Development and demonstration of multidimensional IRT-based internal measures of differential functioning of items and tests. *Journal of Educational Measurement, 34*, 253-272.

Penfield, R. D., & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice, 19*, 5-15.

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

19, 23-37.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19*, 353-368.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*, 552-566.

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

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph, No. 17*.

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

Stocking, M.L., & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201-210.

Stroud, A. H., & Sechrest, D. (1966). *Gaussian quadrature formulas*. New York: Prentice Hall.

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

Teresi, J. A., & Fleishman, J. A. (2007). Differential item functioning and health assessment. *Quality of Life Research, 16*, Supplement 1, 33-42.

Thissen, D.(2003). MULTILOG. In M. du Toit(Ed.), *IRT from SSI*(pp.345-409). Lincolnwood, IL:Scientific Software International, Inc.

Thissen, D., Steinberg, L., & Gerard, M. (1986). Beyond mean group difference: The concept of item bias. *Psychological Bulletin, 99*, 118-128.

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modelling as a Unitary Framework for Binary and Likert-type (Ordinal) Item Scores*. Directorate of Human Resources Research and Evaluation, Department of National Defense: Ottawa, Ont.

Zwick, R., Donoghnue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*, 233-251.

Table 1
Reference Group Item Parameters

Item i	a_i	b_{i2}	b_{i3}	b_{i4}	Item i	a_i	b_{i2}	b_{i3}	b_{i4}
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

Item i	a_i	b_{i2}	b_{i3}	b_{i4}
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

Note. Items not listed used the same item parameters as the reference group.

Table 3
Simulation Design

300 examinees / group	Equivalent ($d_\theta = 0$)	Null Condition (0% DIF)
		Condition 1 (7% DIF)
		Condition 2 (10% DIF)
		Condition 3 (20% DIF)
	Non-Equivalent ($d_\theta = 1$)	Null Condition (0% DIF)
		Condition 1 (7% DIF)
		Condition 2 (10% DIF)
		Condition 3 (20% DIF)
1000 examinees / group	Equivalent ($d_\theta = 0$)	Null Condition (0% DIF)
		Condition 1 (7% DIF)
		Condition 2 (10% DIF)
		Condition 3 (20% DIF)
	Non-Equivalent ($d_\theta = 1$)	Null Condition (0% DIF)
		Condition 1 (7% DIF)
		Condition 2 (10% DIF)
		Condition 3 (20% DIF)

Table 4
Results of LogR (Number of Rejections Out of 100 Trials) for Sample Sizes of 300 and 1000 under the 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. LogR = Logistic Regression Procedure.

Table 5

Results of LRT (Number of Rejections Out of 100 Trials) for Sample Sizes of 300 and 1000 under the 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. LRT = Likelihood Ratio Test.

Table 6

Results of DFIT (Number of Rejections Out of 100 Trials) for Sample Sizes of 300 and 1000 under the 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 Items and Tests Procedure.

Table 7
Mean type I error rates of LogR, LRT, and DFIT (%) under $\alpha = 0.05$.

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

Note.
LogR = Logistic Regression Procedure;
LRT = Likelihood Ratio Test;
DFIT = Differential Functioning Items and Tests Procedure.

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

		$d_\theta = 0$						$d_\theta = 1$					
% of		300			1000			300			1000		
DIF	Item	LogR	LRT	DFIT	LogR	LRT	DFIT	LogR	LRT	DFIT	LogR	LRT	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.
LogR = Logistic Regression Procedure;
LRT = Likelihood Ratio Test;
DFIT = Differential Functioning of Items and Tests Procedure.

Table 9
Mean True Positive Rates (Power)

	LogR	LRT	DFIT
Sample Size			
300	0.1922	0.8268	0.6677
1000	0.3342	0.9833	0.9737
Ability Distribution			
Equal ($d_\theta = 0$)	0.3830	0.9093	0.7840
Unequal ($d_\theta = 1$)	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.
LogR = Logistic Regression Procedure;
LRT = Likelihood Ratio Test;
DFIT = Differential Functioning of Items and Tests Procedure.

Table 10
Mean False Positive Rates

	LogR	LRT	DFIT
Sample Size			
300	0.0055	0.0679	0.0566
1000	0.0151	0.0974	0.0795
Ability Distribution			
Equal ($d_\theta = 0$)	0.0019	0.0798	0.0699
Unequal ($d_\theta = 1$)	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.
LogR = Logistic Regression Procedure;
LRT = Likelihood Ratio Test;
DFIT = Differential Functioning of Items and Tests Procedure.